DOCUMENT RESUME

ED 152 706 TM 006 831

AUTHOR Stake, Robert E., Ed.; And Others
TITLE Case Studies in the Evaluation of Educational Programmes.
INSTITUTION Organisation for Economic Cooperation and Development, Paris (France). Centre for Educational Research and Innovation.
PUB DATE [74]
NOTE 126p.; For related document, see ED 142 565

EDRS PRICE MF-$0.83 HC-$7.35 Plus Postage.
DESCRIPTORS *Case Studies; Compensatory Education; Computer Assisted Instruction; Curriculum Evaluation; Educational Assessment; *Educational Innovation; *Educational Programs; Educational Television; Elementary Secondary Education; Equal Education; Evaluation Methods; Foreign Countries; Gifted; Government Role; Media Technology; National Programs; Post Secondary Education; *Problems; *Program Evaluation; Research and Development Centers; Staff Utilization; State Programs
IDENTIFIERS Canada; Great Britain; Sweden; United States; West Germany

ABSTRACT

Ten brief case studies illustrate the types of arrangements which have been made between commissions and evaluators, and describe some of the subsequent problems. Each of the cases is a major evaluative study of a large-scale educational development or school operation. For the collection as a whole, the purpose was to illustrate how the evaluation got started and how it resolved problems. The case studies are: (1) The UNCAL evaluation of computer assisted learning, (2) the evaluation of the Illinois program for gifted children, (3) the Coleman Report on equality of educational opportunity, (4) the Michigan educational assessment program, (5) the case study of information retrieval television used to access instructional video tapes and films, (6) the use of evaluation as a basis for school reform in Sweden, (7) the assessment of the programs of educational laboratories and research and development centers in the United States, (8) the Biological Sciences Curriculum Study (BSCS) population genetics curriculum field trial, (9) the Project Follow Through evaluation, and the German adaptation of the American Social Science Laboratory Units. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

centre
for
educational
research
and
innovation

CASE STUDIES IN THE EVALUATION OF EDUCATIONAL PROGRAMMES

TM006 831

OECD

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT

CASE STUDIES IN THE EVALUATION OF EDUCATIONAL PROGRAMMES

Collected and edited by

Prof. Robert E. STAKE

of the University of Illinois at Urbana-Champaign

with the collaboration of

Kathleen BROPHY
Darrel N. CAULLEY
Judy DAWSON
J.L. HEINMILLER
Stephen KEMMIS
Marilyn MARTIN
Monika MEINHOLD
Deborah RUGGE
Karin SITTE
Richard SUDWEEKS
Patrick L. WALSH
Eric WEIR
Robert L. WOLF

TABLE OF CONTENTS

I. INTRODUCTION 5

II. THE CASE STUDIES

1. The UNCAL Evaluation of Computer-Assisted Learning 7

2. Case Study of the Evaluation of the Illinois Program for Gifted Children 14

3. Case Study of Equality of Educational Opportunity 22

4. The Michigan Educational Assessment Program 47

5. A Case Study of Information Retrieval Televisica 56

6. Evaluation as a Base for School Reform - The Case of SIA in Sweden 69

7. Assessment of Programmes of Educational Laboratories and Research & Development Centres in the United States, 1972 75

8. An Experimental Approach to Curriculum Evaluation: The BSCS Population Genetics Field Trial 87

9. Case Study of the Evaluation of Project Follow Through 102

10. Case Study of the German Adaptation of the American "Social Science Laboratory Units".....112

III. SUGGESTED READING LIST.....125

Chapter I

INTRODUCTION

Ten brief case studies were prepared by persons affiliated with the Center for Instructional Research and Curriculum Evaluation at the University of Illinois at Urbana-Champaign, to illustrate the different ways in which arrangements have been made between commissioners and evaluators and to describe some of the subsequent problems. These case studies are presented in this document, which should be read in conjunction with the OECD publication "Evaluating Educational Programmes. The Need and the Response", by Professor Robert Stake of the University of Illinois at Urbana-Champaign.

Each of the cases is a major evaluative study of a large-scale educational development or school operation. The ten were chosen partly because of their illustrative value, the accessibility of the records, and an interest in the case by one of the researchers.

When the Steering Group for the CERI project on "New Approaches to Evaluation" met in Paris during the summer of 1974, there was substantial support for making available case studies. There was also a strong feeling that the case studies should be chosen on specific criteria, partly so that they would address the key issues of commissioning and evaluation methodology. Those of us who met on that occasion were unable to agree upon criteria, at least not with assurance that cases fitting the specifications could be found and that case-study writers could be persuaded to study that selection.

Some of the criteria considered were the following :

a) that the programmes and evaluations, for the most part, would be "over and done", and thus studied in retrospect;

b) that the programmes would have been innovative and representative of innovative programmes generally;

c) that the evaluations would have been done by competent researchers;

d) that examples would be drawn from several nations, certainly not all from the United States;

e) that national or state governments or a private funding agency would have been involved in most of the programmes;

f) that some evaluations studied would have been specific to a particular instructional programme but that others would not;

g) that some evaluations would have been oriented to a single school district or within a single district, but that most would have been programmes of national scope;

h) that some of the studies would have been conducted by outside contractors or panels, but that some would have been conducted "in house";

i) that some of the evaluation studies would have taken a strong psychometric or even experimentalist approach, but that some would be cases where other approaches had been taken;

j) that some of the cases would have been curriculum-development projects;

k) that different levels of education - elementary, secondary, tertiary, and continuing - would be included;

l) that some of the evaluative studies would be seen as successful, and others included would not.

For the collection as a whole the purpose was to illustrate how the evaluation got started and how it resolved problems during its lifetime. A list of about twenty evaluated programmes, which roughly conformed to these criteria, was prepared. Faculty and graduate students were each invited to select one of the cases, to work collaboratively and under some direction so as to increase the uniformity and common purpose of the effort, and to prepare a report on each of no more than 2,500 words. The length was limited in order to stay within the anticipated page limits of this document and to increase the likelihood of having it available for the OECD-CERI Conference on Evaluation at Liège, Belgium, in December 1974.

None of the case studies is without flaw. Partly because they are so brief some are oversimplified. In several cases the draft copies circulated to participants drew such objection and reproach that the writer was tempted to abandon the effort but was persuaded then to continue. In some cases the difference in choice of language, perspective, or valuing was resolved; but in some cases (the BSCS Field Study and the Project Follow Through evaluations, to name two) strong disagreement as to what happened remainded as we gathered final copy for this document. The authors (including Robert Stake) were convinced that the accuracy and detail is sufficient to justify close attention to them for instructional purposes. They should not be considered as audits or definitive statements of the evaluation projects, but the issues they suggest are real and commonplace.

Chapter II

THE CASE STUDIES

1. THE UNCAL EVALUATION OF COMPUTER-ASSISTED LEARNING

In December 1967 Britain's National Council for Educational Technology (NCET), later to become the Council for Educational Technology for the United Kingdom (CET), set up a Working Party to "investigate the potential role of the computer as a component of educational and training systems in the United Kingdom taking into account as necessary experience and trends in other countries", and "to outline a systematic programme of applied research and development which it would be desirable to encourage in (the United Kingdom); aimed at exploiting the computer to the best advantage in education and training".(1) On the basis of the report of this Working Party, and a major feasibility study, the National Development Programme in Computer-Assisted Learning (NDPCAL) was created.

On January 1, 1973, NDPCAL started work under its Director, Richard Hopper, previously a Senior BBC Producer in the Faculty of Educational Studies at the Open University. It was initially funded with a government grant of £ 2 million for five years, to operate under the guidance of a Programme Committee which includes representatives of the seven government departments sponsoring it.(2) The Programme Director formulated the aim of the National Programme and

---

1) Computers for Education, Working Paper 1, National Council for Educational Technology, 1969. Two other documents served as useful resources for this case study, both authored by Richard Hooper: "The National Development Programme in Computer-Assisted Learning", CET Information, July 1973 and "The National Development Programme in Computer-Assisted Learning - Origins and Starting Point", Programmed Learning, March 1974.

2) Department of Education and Science, Scottish Education Department, Welsh Education Office, Ministry of Education for Northern Ireland, Ministry of Defense, Department of Employment, and Department of Industry. The Committee also includes representatives from the Council for Educational Technology for the United Kingdom, the Universities Grants Committee, the Schools Council, and the Social Science Research Council.

determined criteria for the design and selection of development projects to meet this aim. This policy was endorsed by the Programme Committee. The projects were designed to be relatively autonomous from the NDPCAL Directorate, sponsored jointly by the educational institutions in which they were to function and by NDPCAL which would provide matching funds. The purpose of this matched-granting arrangement was to increase the chances that projects would continue beyond the term of NDPCAL funding; they were expected to take root in those institutions and to develop into transferable form, that is, a form that could be implemented elsewhere. From the start, NDPCAL had focused some of its attention on assimilation: finding ways to have the innovative computer-assisted learning projects take root in their own institutions and in other institutions, within the normal contexts of those adopting agencies.

The individual projects cover a range of educational settings, but all are concerned with use of the computer as a tool for, and manager of, learning. The major focal settings are: mathematics, science and medicine in upper and post-secondary education; computer-managed learning in the school sector; and industrial training in the armed services and civilian sector. Three examples may suffice to illustrate the diversity:

1. The "Computer-based Laboratory for Service Mathematics Teaching" project, operating at Napier College of Commerce and Technology in Edinburgh and Falkirk Technical College, is concerned with using computers in "service mathematics" courses. These courses are ones in which the mathematics taught is in support of a main science, engineering, business or management subject. The courses therefore focus on the description and use of mathematical systems in practical contexts without requiring detailed knowledge of the mathematical processes involved. The computer takes over the burden of analytical or numerical solution in solving problems once they have been modelled in mathematical terms.

2. At University College, London, Chelsea College, London, and the University of Surrey, the "Computers in the Undergraduate Science Curriculum" project provides computer possibilities for teaching and learning in several sciences, with emphasis on the presentation of graphics. The computers are used for simulation, modelling, data handling and statistical analysis, and for making calculations too tedious to be done by hand. Educational technologists work with teachers and students in physics, chemistry, physiology and biology in developing self-contained instructional packages to cover small and defined areas of syllabus, with an eye to maximizing transferability across institutions and departments.

3. The Hertfordshire Computer-Managed Mathematics project is conducted under the auspices of the Advisory Unit for Computer-Based Education in the Hertfordshire Education Authority. The computer is used in the first two years of secondary mathematics (with 11 and 12 year olds). The computer is used primarily for the management of student responses.

The evaluation components of the NDPCAL Programme were built in from the start. Each of the projects funded by NDPCAL was to have its own evaluation capacity, and NDPCAL was to develop an overarching evaluation capacity of its own. This overarching evaluation capacity was to have two major thrusts: within the Directorate, two staff members were given separate evaluation responsibilities for the educational and computer/technical aspects of the Programme. Outside the Directorate, two independent evaluation efforts were mounted, one fiscal and one on programme development. The latter independent effort is of primary concern here. Further evaluation responsibility was shared, of course, by the Programme Committee. An early issue confronted by Hooper and the Programme Committee was: "Is it too early to have evaluation before you have projects?"

As early as January of 1973, Hooper consulted Barry MacDonald of the Centre for Applied Research in Education of the University of East Anglia about independent evaluations. Previously MacDonald had developed some issue-based collections of testimonial data for the Humanities Curriculum Project, a data type suiting Hooper's preferences for "illuminative evaluation". Hooper drafted a plan for a three-year study and proposed to the Programme Committee that it be an independent study directed by MacDonald. MacDonald was interviewed by a subcommittee of three Programme Committee members plus a civil-service research administrator and a professor of psychology.

In a formal session several issues were discussed at length:

a. what to do in case evaluator and steering committee disagree about design or conduct of the study,

b. the relative importance of observational and psychometric measurement data,

c. the importance of the evaluator being an expert in the particular subject matter and technology of instruction used,

d. whether evaluation in this case should be for the purpose of providing information to facilitate decision-making or to justify it, and

e. whether cost evaluation and substantive evaluation should be combined or separate.

Some members were unfamiliar with the evaluation methods MacDonald used and no one was certain as to what technique would be most needed. The subcommittee deliberated long, then authorized Hooper to sign MacDonald as a consultant for six months, June through November.

MacDonald thus had an opportunity to examine individual projects and the Programme as a whole, and to plan the three-year study more carefully. The Programme Committee later voted to support his plan starting in January 1974.

The group to conduct the evaluation was led by MacDonald, budgeted for two-fifths of his time. David Jenkins and David Tawney began work as full-time research associates in April (1974). A fourth evaluator, Stephen Kemmis, is to be added for 1975 and 1976.

The UNCAL [1] evaluation study, as MacDonald's programme-development evaluation of NDPCAL came to be known, is by comparison with many evaluation studies a very large project, having a budget of some £91,000 for three years. MacDonald's proposal stressed four particular evaluation roles :

a. a documentation role, producing a full narrative history of the Programme,

b. a responsive role, providing information to help meet the information needs of people within and outside the Programme, especially decision-makers,

c. a developmental role, offering evaluation advice and consultancy to individual projects, and

d. a communication/dissemination role, displaying information to mediate between producers and consumers of ideas about the developments of the Programme.

The UNCAL evaluation can therefore be seen as operating alongside, but independently of, NDPCAL's other evaluation efforts, with major responsibilities for documentation and for facilitating the exchange of information among various groups within NDPCAL, i.e., projects, Directorate, Programme Committee.

The history of NDPCAL prior to the start of UNCAL evaluation shows an early and deliberate review of many methodological issues. Among these issues :

a. the definition of evaluation to be selected,

b. the goals of the evaluation,

c. the range of responsibilities of the evaluators,

d. the terms of cooperation between individual project evaluation efforts and the UNCAL group, including rights of access to the individual projects by UNCAL, and

e. the evaluation methodology and the particular tasks of each evaluation group.

Several of these issues were seen to have ethical as well as methodological implications.

In spite of all this early deliberation, challenges continue to be raised about fundamental aspects of UNCAL's role : whether it has responsibility for evaluation of the Programme Directorate's work, what data are appropriate in feedback to projects, what questions are

---

1) Acronym for Understanding Computer-Assisted Learning.

to be addressed by the evaluation, and whether it has "generalisable-research" responsibilities to confront the basic questions about computer-assisted learning. The aims of Hooper and MacDonald on these matters are on record, but even between them there is less than a perfect understanding. Even MacDonald's definition of evaluation has been challenged :

> "Evaluation is the process of conceiving, obtaining and communicating information for the guidance of educational decision-making with regard to a specified programme."

It is a definition clearly different from the student-measurement approach often taken in evaluation, unless one assumes that national programme decisions can and should wait until student achievement data are available.

From the beginning the UNCAL evaluation team has appeared to make its programme monitoring role the basis for its other roles. Its first semi-formal report, presented after three months, took the form of a dialogue between two fictitious persons. It raised issues that had been expressed to the evaluators in the course of their monitoring visits. Sometimes the issues were stated in the actual words of informants as taken from transcripts of taped interviews, although quoted in the report without attribution. It had been hoped that the dialogue format and the identification, documentation and presentation of "issues as experienced" would capture some of the complex reality of the Programme, including its political side, and would provoke further consideration of these issues by the Committee and the Directorate.

Some of the issues discussed in the Issues Report included :

- matters related to expectations of the Programme created by the large-scale funding, e.g., that there is a tendency to portray it as a sound investment,
- questions about the political context in which NDPCAL and the Programme Committee operate,
- problems of central administration of projects that are expected and expecting to be autonomous,
- allocation of sufficient attention to the matter of transferring what is developed in one college to another,
- ambiguities in the relation between research and development for the projects, e.g., the extent to which development opportunities are being exploited at the expense of research,
- problems in the evaluation itself, e.g., the extent to which there is pressure on project evaluators to become advocates for their projects, and
- problems of the degree to which Programme administration and documentation absorbs the energies of the project staff.

In as highly political a context as the one in which NDPCAL exists, seen as it is as a major, innovative, highly visible, model programme, it is not surprising that issues arise again and again like phoenixes from the ashes of past decisions.

Circulation of the Issues Report to the Directorate and Programme Committee did not precipitate a new look at the issues - it precipitated a strong protest about UNCAL's approach to evaluation. Several members of the Programme Committee were annoyed with MacDonald for departing so far from "accepted" methods of research and evaluation. Even after MacDonald reviewed the several ways in which his team was carrying out the details of his proposal, there were objections to the continuation of the contract. The Committee decided to keep the contract, but MacDonald was reminded again of "what decision-makers need to know," particularly more about how student learning is affected and particularly less about what an occasional critic says about the Programme.

The misgivings expressed within the Committee after the presentation of the Issues Report had not been dispelled by the restatement of the same issues in a different format, not a dialogue but a formal listing. The new report seemed to be ignored. Apparently the reporting of issues itself, or the time spent in collecting the issues, was in question. Those in protest, including several who had participated in the subcommittee interview of MacDonald, wanted answers to questions, not reminders of questions, even at that early stage.

In spite of UNCAL's attempt to be faithful to complexity (or, more probably, because of it) Committee members were left with the nagging doubt that their information needs will not be met.

Perhaps the twin concerns for good documentation ("telling the story" of the Programme faithfully) and for responsiveness to the information needs of the decision-maker are diametrically opposed here, rather than in dialectical opposition. The problem seems to be a failure of communication rather than of accurate representation. Expectations of research-oriented reports focusing on student achievement and control-group experiments have not been met. The failure of communication here is symptomatic of a conflict between different views of evaluation ("different ideologies", as one party to the scene construes it).

All this happened in the first half-year of a three-year study. Late this same period two other UNCAL documents were circulated to the Directorate and on to the Programme Committee, one on the several roles of UNCAL and the other on David Tawney's consultative work with projects developing science courses.

It may be of no significance but it has been noted that UNCAL does not feature visual representations, as some evaluation studies do. With a preference for written and verbal reporting UNCAL may fail to satisfy some audience needs for models, charts, and other visual materials.

The issue of how UNCAL can provide support to the projects remains partly in question. Direct personal aid has been successful in some projects but rapidly exhausts staff time, and is just not appreciated in some quarters. Early efforts to circulate general information on how to do formative evaluation and case study information on the evaluation of computer-aided instruction elsewhere were not fruitful. Most of the writings found were objectionable for one reason or other. The "jargon barrier" was found to be difficult to penetrate. Even when doubly cautious the researcher is unable to avoid using the special language of his craft.

It seems that UNCAL has, at least partly successfully, avoided the "centre-spy" role, one certain to jeopardize relations with projects. In this early period there appeared to be little apprehension in the projects (or on the part of the Directorate) about UNCAL. Hooper expects UNCAL to counsel each project specifically so that it can make the best possible case for itself when annual funding decisions are again to be made. With one or two exceptions, UNCAL's relations with the projects seem suitable for this task.

In general, the UNCAL evaluation study is characterised by great attention to design and execution of evaluation activities. Yet in spite of conscientious work, explicit plans, and thorough negotiations by experienced professional people, a disagreement remains. Commissioners and evaluators might have been able to work out these differences prior to writing a contract, but they may have done all they could do at that time. The needs for data, and the opportunities for gathering it (as at least one Programme Committee member foresaw) would evolve as the national programme evolved. The disagreement became more acutely apparent because of the self-imposed obligation on UNCAL for providing frequent reports of programme progress and hazards.

Two possible - but speculative - reasons for the methodological disagreement are first, that MacDonald and some Programme Committee members disagree about the ideological nature of evaluation and, second, that the UNCAL staff is interested in evaluating the world of evaluation at the same time it is evaluating the National Developmental Programme for Computer-Assisted Learning. Many evaluation studies suffer from the false presumption that the procedures used are the only proper procedures possible. This evaluation study may suffer from a fascination with itself and its role in explaining the multiplicity of evaluation methods.

At this time, the UNCAL evaluation study is still in progress. It will continue, according to the contract, until December of 1976.

2. CASE STUDY OF THE EVALUATION OF THE ILLINOIS PROGRAM FOR GIFTED CHILDREN

One of the first large-scale independent evaluation studies in the U.S.A. was of the Illinois State Program for Gifted Children. The "Gifted Program" itself was preceded by a four-year study that found special provision for education of the gifted child nonexistent in the state. When the Program was proposed to the state legislature in 1963, the proposal included the recommendation that the Program, as an experiment, should be evaluated after the first few years. Large in size and budget, the Gifted Program anticipated large-scale educational change in the state. Its proponents wanted an evaluation to determine whether or not the Program made a difference to mentally gifted children in Illinois.

The evaluation of the Gifted Program, from initial planning to last report, ran four years. This evaluation came at the beginning of the evaluation/accountability movement in the U.S. It was more or less without precedent. Few models were available. Both the evaluator and the commissioners of the evaluation were making fresh tracks rather than following a well-worn path.

The evaluation began in 1967 and was completed in 1971. Time has shown both the statewide Program and its large-scale evaluation to be good efforts. Time also has dimmed the memory of pains involved with the Program's early operation and its evaluation. Only major events remain imprinted in the minds of the participants. Of those interviewed, however, each tell amazingly consistent tales of the initiation of the evaluation, although from definitely different perspectives.

The main characters on the scene at the time of the evaluation were :

David Jackson - one of the fathers of the Gifted Program, first director of it, and the instigator of the evaluation of the Program;

Ernest House - the evaluator in charge of the evaluation;

Herbert Baker - second director of the Gifted Program, in charge when the evaluation was planned and begun;

Robert Hardy - third director of the Gifted Program, in charge when the last evaluation report was submitted;

OSPI - the Office of the Superintendent of Public Instruction of the state of Illinois, within which the Gifted Program operated; this office also funded the evaluation of the Program.

The Gifted Program began in 1963, first headed by Jackson. The Program's efforts toward educational change in the state were several. Its money financed five components : administration of the Program from OSPI offices; demonstration centers which ran model classrooms for teachers to observe; experimental projects to find new ways of teaching the gifted; reimbursements to public schools with programmes for the gifted; as well as training programmes for teachers and administrators. Each school defined - however it wished - who was gifted and what the local gifted project should entail. Projects ranged from music, dramatics and creativity programmes to accelerated classes in the "three R's". Such variability and flexibility made the state Program a challenge to evaluate.

Although Jackson had deep commitments to the need for evaluating programmes in order to justify their existence, the first two funding biennia passed without formal evaluation except for review of the ledger sheets (i.e. cost per child, number of classrooms involved,etc.). Jackson said later that he "failed" by not building evaluation into the Program while he was its first director. According to him: "One of the most pressing needs of modern education is to document and evaluate past programs in order to lay the foundation for future education."

After serving as its first director Jackson served on the Advisory Council to the state staff of the Gifted Program. Several of this Council felt that evaluation of the Program was imperative. With their encouragement Jackson submitted more than one plan for evaluation to the state staff. However, the state staff (some 12 civil servants), unfamiliar with comprehensive evaluation of large programmes, felt that the plans were extravagant, involving too much time, staff and money. They had higher priority items for the money that had been allocated. Sometime later (1967) it became apparent to the staff that many of the "experimental programmes" were not producing what had been anticipated, and further that substantial research money was paid to research assistants on university campuses - persons who seemed to think of the Gifted Program as a way to get their way paid through graduate school. Following this disillusionment it became possible to persuade the state staff that some of the money originally budgeted for experimental projects could be budgeted for an evaluation.

Jackson, now heading a federally funded Regional Research and Development Laboratory (CERLI) in Illinois, worked carefully to sell the Council's idea of an evaluation of the Gifted Program. He went to the man in charge of the OSPI office, the State School Superintendent, Ray Page. Then he worked his way down the line of assistant superintendents, arguing his case, and then "reversed the process", up the

hierarchy, talking finally again with Page.[1] These meetings culminated with a large luncheon, bringing all of them together. Jackson assured them that the main focus of the evaluation would not be the administration of the Gifted Program. He received permission to go ahead with an initial study, preparatory to doing the evaluation. Jackson thought he finally had the evaluation sold.

Jackson then hired Ernest House as the evaluator upon the recommendation of his academic advisor, James Gallagher, one of the researchers who had conducted the original study of opportunities for gifted children in the state. Jackson and House wrote a brief (3 page) outline of how they would go about studying the Gifted Program in Illinois for a year in order to produce an evaluation plan. Page supported the plan of action, but requested in addition a survey and data to be reported out in eight months since the Gifted Program itself would need support the next year in the legislature.

House worked for the next few months, drawing advice from a panel of evaluation experts and from written evaluation theories and models. He drew up a plan for three years of evaluation. His plan was a general outline, not a tight design - the plan could be modified and revised as the situation demanded. It was apparent that standardized achievement tests were not appropriate for this assignment. Extensive observation, questionnaires and interviews were needed to gather evidence and to create a picture of the Gifted Program.

In the spring (of 1968) he hired staff to do a statewide mail survey to provide the Superintendent with evidence of support of the Gifted Program. A report was written by House and sent to the Gifted staff, with sections underlined in red by Jackson to indicate potential areas of action. Jackson and House were ready to go with the three years of evaluation.

However, OSPI was not. A new Associate Superintendent, Robert Cain, had come into OSPI to head the section in which the Gifted Program resided. He saw that a great deal of money ($400,000) was being allocated for an evaluation study that he had little control over. Talking to members of the state staff in his office, it became apparent that Cain was worried that Jackson was being allowed more administrative control over the Program than an outsider should have. The report from House "gathered dust".

---

(1) Jackson recognized that the choice of state programmes to be funded is not entirely a political decision - but with more good programmes than available funds can support, it is necessary for an educator to be sophisticated in presenting information that will make administrators and legislators aware of the worth and needs of a project.

The next move was Jackson's. Both the Gifted Program and its evaluation now were in great danger of losing their funding, and Jackson was a strong advocate of both. He asked for and received a hearing before the School Problems Commission, a legislative/public committee which makes recommendations to the state legislature as to what programmes ought to be funded during the coming biennium. In addition he and House "scurried around" talking to OSPI officials and legislators to assure them that Jackson was not trying to "run the show" and that the state would get a disinterested, comprehensive evaluation of the Gifted Program. Although several people thought Jackson's actions were "as much a part of the problem as the solution", the lobbying worked.

The CERLI Lab got the money for House to continue a second year of evaluation. Funding had to be negotiated during each of the next two years also, although with considerably less opposition.

Politics were a major stumbling block in the negotiations of the evaluation of the Gifted Program. Perhaps it was inevitable. The money and personalities involved were substantial. Money and its politics also influenced the way in which the evaluation study was executed.

The funds for the evaluation of the Gifted Program came out of the OSPI annual budget. For the four year evaluation of the original plan was budgeted at about $ 100,000 per year (out of a total Program budget of $ 8,000,000). The money was to be allocated at different rates over the four years. The costs were expected to vary, lightest for the planning year, heaviest for the data collection, not quite so heavy for the analysis and report-writing years. House decided after the first year that this might not be the best use of the money for the evaluation, but Jackson suggested that it would be best not "to rock the boat" by asking for a different allocation pattern.

House would have preferred an annual allocation based on the evaluation plan rather than an evaluation plan based on the allocations. He thought it would have been better to have had yearly cycles of planning, data collection, analysis and write-up.

At the end of the second year of evaluation it became apparent to people advising House in his evaluation that the CERLI Lab was to be terminated. House got a new group, CIRCE, to sponsor the evaluation. Nothing was changed in the evaluation operational plan, but the headquarters for the study shifted to a different community - the community of his advisory group, and of his university colleagues - one that allowed greater access to the state staff.

House and his staff brought with them to their new offices on the University of Illinois campus all the data they had gathered the prior year. There was a great amount of it, and it needed to be analyzed.

At this point House would have preferred to have been out gathering some different kinds of evidence of programme success but felt committed to analyze the data already gathered.

The evaluation master plan seemed to suggest that the product of the evaluation would be one large report at the end of the four years. However, thirteen reports were written and distributed over the four years. There were four reports the first year, three the next, four the next and then two the last. These were reports intended for a public audience (i.e. legislators) not for confidential feedback to the Program developers or educators in the local projects. In addition there were reports to the individual demonstration centres, providing staffs with information about their own centres as compared to others.

The evaluation was being done primarily for those who would determine the fate of the Program, but the Program staff and hundreds of teachers in the field were recipients of information also. Some of it helped them make decisions about the day to day and month to month direction of their efforts. This "spinoff" became increasingly important, even though it was not the original intent of those who authorized the evaluation.

All of the people involved with the Illinois Program for Gifted Children and with its evaluation maintain that several major changes were made in the implementation of the Program and of the strategy for educational change as a result of the evaluation. House would have preferred that the state staff, upon seeing and hearing the information that he and his team had gathered, would draw their own conclusions and make their own decisions about changes necessary in the Program. However, at the final presentation of data session Associate Superintendent Cain rejected conditional statements telling the evaluators in essence that they should not accept so much money for doing an evaluation if they are unwilling to use their own data and judgment to formulate recommendations. Therefore, against his better judgment, House drew up a list of recommended changes for the Program. Many later were implemented. About the Program House concluded:

> "While it discovered some of the premises on which it was founded not to be true, it has subjected itself to intensive critical examination and has taken broad adjustments on that basis. The final judgment then is that the Illinois Plan has given the state far more than its money's worth."

Hindsight

In general all of the major participants in the initiation of the four-year evaluation of the Illinois State Program for Gifted Children now (1974) feel satisfied with what was done. But there are, of course, ideas of how to proceed differently "the next time". They now have been through at least one major evaluation (and in some cases several) and the field of evaluation and the issues surrounding it have proliferated in the U.S. since 1971 when the final report was submitted.

Baker, the director of the Illinois Gifted Program for most of the four-year evaluation, commented that House was in touch with his (Baker's) staff about every two weeks. He said that House maintained excellent contact. Some of the contact was needed to keep the evaluators in touch with Program developments, and some to assuage the feelings of the Gifted staff. They could talk with House, know what he was doing, and where. The contact was not to get instructions from the state Gifted administrators. This staff and their Advisory Council wanted an evaluation of the Program to let them know if the Program made a difference in the state; they thought it did, but they wanted an independent documentation. In spite of their commitment to the evaluation they did not know what it could do for them nor how it could be done. They expected the evaluators to define the issues and to decide when, where, and how the data would be collected. But now, these education officials - having become more sophisticated about evaluation - feel that they would be more definitive as to what the evaluators should be doing for them.

Hardy, the subsequent director, and House, the evaluator, expressed thoughts about what they would discuss if a new evaluation were being contracted. Hardy said that the commissioner of the evaluation should "Detail 1, 2, 3, 4..." each of those things the evaluator should look at and be answering in the evaluation. The issues he then said he would like examined are almost those of an educational research effort. He would like information from the evaluation on "how to identify the gifted and talented child" and "how to individualize instruction for them". It is to be recalled that the evaluation study got funded only when experimental studies with such aims failed, thus releasing funds. House at the same time felt that prior to beginning an evaluation there should be full discussion of the situation, the criteria for the evaluation and the weaknesses of research methodologies for providing evaluation information.

Jackson, who spent several years trying to get the evaluation of the project started, more recently has thoughts about how to ensure that the evaluation gets done. He suggested that by deliberately involving many persons critical to the survival of a project in the

plans for the evaluation that an expectancy for data can be built up that in effect ensures that the evaluation will be done.

An evaluation study appears to be a product of the social climate of the times. Issues not considered critical by the evaluation's audiences most likely will not be pursued and documented. For example, the Gifted Program had wanted to know if its components separately, and the components together as a whole programme, were doing what they had been designed to do. The standards against which to judge this Program were drawn from its design. An issue the evaluators felt important was whether an "elitism" might be fostered by the Program which would be inconsistent with the goals of a democratic society. No audience of this evaluation called for answers to this question. Currently, in 1974, the Gifted Program wants a new long term comprehensive evaluation. The social climate will probably demand that external political standards as well as internal bureaucratic standards be recognized again, if another evaluation study is to take place.

REFERENCES/SOURCES

BAKER, Herbert, Second Director of the Illinois State Program for Gifted Children. Telephone conversations, October and November, 1974.

HARDY, Robert, Third Director of the Illinois State Program for Gifted Children. Telephone conversations, October and November, 1974.

HOUSE, Ernest, Evaluator of the Illinois State Program for Gifted Children. Personal conversations, October and November, 1974.

JACKSON, David, Director of the Coordinated Educational Research Laboratory and First Director of the Illinois State Program for Gifted Children. Telephone conversations, October and November, 1974.

Three of the evaluation reports provide a good overview of the Gifted Program and its evaluation:

HOUSE, Ernest R., Kerins, Thomas and Steele, Joe M. The Demonstration Center. December 1970.

ILLINOIS Gifted Program Final Evaluation Report. Educating the Talented. June 1971.

STEELE, Joe M., House, Ernest R., Lapun, Stephen, Kerins, Thomas. The Gifted Classroom. August 1970.

There are a limited number of copies available from:

Gifted Children's Section
The Illinois Office of Education
Department of Exceptional Children
1020 South Spring Street
Springfield, Illinois 62706
U.S.A.

3. CASE STUDY OF EQUALITY OF EDUCATIONAL OPPORTUNITY

In 1966, the report "Equality of Educational Opportunity"[1] was submitted to the President and Congress of the United States of America by the then U.S. Commissioner of Education,[2] Harold Howe II. The Commissioner had conducted the survey as required by Section 402 of the Civil Rights Act of 1964. James Coleman of Johns Hopkins University had major responsibility for the design, administration, and analysis of the survey. The report came to be referred to as the Coleman Report and will be termed such in this case study.

Origins and Objectives

Section 402 of the Civil Rights Act of 1964 passed by the Congress of the United States of America states:

> Sec. 402. The Commissioner shall conduct a survey and make a report to the President and the Congress, within two years of the enactment of this title, concerning the lack of availability of equal educational opportunities for individuals by reason of race, color, religion, or national origin in public educational institutions at all levels in the United States, its territories and possessions, and the District of Columbia.

While this Section represents the explicit commissioning of the survey, commentators agree that its origins and the objectives Congress had for the survey are not very clear. The origins of the survey are related to social, economic and political climate of the time.

> In the course of the 1960's the critique of social myths and beliefs acquired a significant role in American national life. Economists had for some time secured themselves a position in national councils, but now a wide range of other disciplines was brought in and social science information became increasingly central to the formulation and evaluation of social policies... A combination of political, economic, and social events led to unprecedented national commitments to restructure the society so as to overcome the injustices and instabilities associated with the historic problems of poverty and race. Simultaneously the methodology of the social sciences made rapid advances, primarily those associated with the development of the high-speed computer, so that large-scale, complex investigations became possible to a degree that had never previously existed (Mosteller and Moynihan, 1972, pp.3-4).

---

1) James S. Coleman, Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York, Equality of Educational Opportunity. Washington,D.C. U.S. Government Printing Office, 1966.

2) The U.S. Commissioner of Education is in charge of the Office of Education which falls within the Department of Health, Education and Welfare (HEW).

According to Gerald Grant[1] the Civil Rights Act of 1964 was conceived during the Presidency of John Kennedy, languished in the Congress, and then was enacted in the emotional days following Kennedy's assassination partly as a testament to him. The educational aspects of the Act received little notice during its passage through Congress. Attention was focused on the highly controversial aspects of the Act - voting rights and public accommodation.

The bill had originated in the Justice Department. Section 402 "had originally been written in rather punitive language, giving the Commissioner of Education power to 'conduct investigations' of 'the extent to which equal educational opportunities are denied to individuals...'" In last minute compromises in the Justice Department and on Capitol Hill, large sections of the bill were rewritten and toned down to gain support for its passage[2] (Grant, 1973, pp.18-19). The spectre of federal education officials conducting investigations in racially sensitive areas of southern United States was neatly replaced by the image of just another survey.

> If any clearly conceived objective for the survey did exist, the sequence of events obscured and muddled it. Initially Congress seems to have intended the study to become a tool for legal actions opposing formal discrimination against minority groups. As it became evident that the statute would forbid outright any such discriminatory acts, the final intent may have been to establish once and for all that gross differences in school facilities did exist, especially as between black and white children in the United States (Mosteller and Moynihan, 1972, p.5).

The first research task of Coleman in carrying out the survey was to "discover" the intent of Congress.

> The first and suprising result was that there had been a shift in the intent of Congress from the initial inclusion of Section 402 to the final passage of the bill. Initially, the section, in different form, was of principal interest to the Justice Department, which probably was responsible for its initiation, and was not a matter of policy research at all. It was instead designed to discover intentional acts of discrimination on the part of school systems to allow for prosecution by the Justice Department. Then as the civil rights bill proceeded through Congress, the Justice Department lost all interest in the section. It would have been dropped had it not been picked out by some Senators and transformed into a policy research question. It was clear from our investigation that the intent of Congress was to discover inequalities in educational opportunity, and sources

---

1) Gerald Grant investigated the politics of the Coleman Report and much of the material for this case study is based on his work which is reported in the article "Shaping Social Policy: The Politics of the Coleman Report", Teachers College Record, Vol. 75, No.1, 1973, pp. 17-56.

2) Gerald Grant's interview with Bernard Zelenko, General Cousel, Committee on the Judiciary, House of Representatives.

of inequalities, as the basis of future policy designed to reduce them.(1) This left, however, a thorny conceptual problem: what was meant by "inequality of educational opportunity". A review of the statements of members of Congress indicated that different Congressmen meant very different things by it (Coleman, 1972, p.7).

In summary, certain points emerge. The first draft of a parliamentary act is often written by the appropriate civil service departments. Such a draft will often reflect the concerns of the civil servants involved, and as in the case of the Civil Rights Act of 1964, their need for information in order to perform their responsibilities. In the course of its passage through the parliamentary body, the act will be modified and changed, by members of parliament and often by members of the relevant civil service departments. If the passage of an act occurs over a considerable period of time the concerns and intents of the participants may change with changing social climate and thought. Behind explicit statements commissioning an evaluative survey or investigation will be unstated intents and purposes which will vary across the individuals which were in some way connected with the commissioning.

The Selection of the Evaluators

The Civil Rights Act of 1964 had originated in the Justice Department and officials in the U.S. Office of Education did not have any clear idea of what to do about Section 402. Perhaps by default, the administration of the section fell to David Seeley, a lawyer and aide to the then Commissioner of Education, Francis Keppel. Seeley farmed out the assignment to lawyers in half a dozen cities who began looking into desegregation problems on a part-time basis (Grant, p.18).

This situation remained in force until Keppel brought Alexander Mood, a former executive of the Rand Corporation, to head the National Center for Educational Statistics within the U.S. Office of Education. As a mathematician, Mood was familiar with the use of statistics and computerized survey techniques.

> One of the first items Keppel and Mood discussed was the "402 survey". Keppel was worried that the scope of Seeley's project was too small. Mood agreed that it was of little value and went to work on a new design. This was a radical step that would involve the agency (the Office of Education) for the first time in the politically sensitive sponsorship of classroom testing.... But he (Keppel) decided to take the risk, and transferred responsibility for the survey from Seeley's office to Mood's... (Grant, p.19).

---

1) Coleman places a footnote here: "It is curious that by an ironic twist, the results of the report have probably been used more by the courts, in the prosecution of cases on school desegragation, than by either the Congress, the executive branch of government, or local school boards".

It was Mood's, and later Coleman's, idea that the survey examine educational achievement and attempt to relate these to the inputs of the educational system. The U.S. Office of Education had traditionally been an inventory-taking agency of inputs. School administrators were adverse to federal testing programmes in the schools. Thus it seemed prudent as well as practically necessary for the Office of Education to bring in outside personnel for the survey.

> Mood, impressed with Coleman's work... and with his ability to handle large-scale survey research, asked the Johns Hopkins sociologist to undertake the major responsibility for the survey... Coleman viewed the appointment of a man of Mood's background as a demonstration of new scholarly intent at the agency (the Office of Education), erasing any doubt he might have formerly had about the quality of research that was envisioned. He accepted the assignment with some pleasure, as co-director with Ernest Q. Campbell of Vanderbilt (University), who took charge of the less publicized higher education aspects of the survey (Grant, pp. 19-20).

Thus the survey was carried out by Mood's office, the National Center for Educational Statistics. Coleman had major responsibility for the design, administration, and analysis of the survey and Campbell shared this responsibility. Staff members of the Center assigned full time to the survey were Carol Hobson, James McPartland, Frederic Weinfeld, and Robert York and these four people, together with Coleman, Campbell and Mood became co-authors of the final report of the survey.

In effect Coleman became a government employee within the U.S. Office of Education. He was in the somewhat ambivalent position of being both project director and a consultant for the study. There was no contractual agreement between Coleman and the Office of Education in relation to the survey. Coleman realized that there were no guarantees with regard to such issues as freedom with respect to choice of methods of inquiry, control over the content of the final report, publication of results in journals and so on. However, Coleman accepted the assignment because of his regard for Mood who he felt would back him on any issues that arose during the conduct and reporting of the survey. Thus it is possible to distinguish between the situation in which a government agency brings an individual into the agency to take charge of a specific study, and the situation in which a government agency lets out a contract to an individual to conduct a study.

The major investigation involved the survey of primary and secondary schools. A number of other studies were carried out and some were quite extensive. Campbell had major responsibility for the higher education surveys. Florida State University was contractor for an extensive study of school enrolment and dropout rates of children of various ages, races and sociometric categories. Raymond W. Mack of Northwestern University directed a team of sociologists who did case studies of education for minorities in ten American cities.

Keppel appointed an eighteen-member advisory committee, half of whom were university professors in the social science disciplines, and the other half were mainly school superintendents of large school districts. The advisory committee assisted in the design of the study and in developing procedures for carrying it out. However, it did not participate in the analysis of the data nor the preparation of the final report.

The Design of the Evaluation

Coleman saw Section 402 as requiring policy research as distinct from discipline research.

> The defining characteristics of policy research are two: the research problem originates outside the discipline, in the world of action; and the research results are destined for the world of action, outside the discipline. The special properties of policy research stem from the different properties of the disciplinary world and the world of action, and from the translation problems involved in moving between these two worlds.(1),.. Specifically, the design of research in social policy should follow these steps:
>
> 1. identify the parties interested in policy outcomes and with some power or potential power to affect policy;
>
> 2. determine interests of these parties;
>
> 3. find what kinds of information are relevant to their interests;
>
> 4. determine the best way to obtain this information;
>
> 5. determine how to report the results (Coleman, 1972, p.16).

Coleman has seen the latter three steps as constituting primarily technical problems of research design concerned with optimum methods for obtaining certain kinds of information and for analysis and reporting the results. Furthermore, these technical problems are largely identical with those in discipline research.

The first task of the survey staff involved carrying out the first two steps. A major difficulty was the determination of what was meant by "inequality of educational opportunity". A study of statements of Congressmen showed a wide disparity of meanings.

> Consultation with members of interest groups (Negro civil rights groups, religious groups, ethnic interest groups, as, for example, Puerto Ricans) gave no narrowing of the range of things that were meant by the term, no convergence on a single generally accepted meaning of "equality of educational opportunity". Finally, an historical examination of the meaning of the concept showed an evolution involving several distinct meanings....But the historical

---

1) James S. Coleman, "Policy Research in the Social Sciences". Morristown, New Jersey: General Learning Corporation, 1972, p.3. Coleman discusses four important properties of the world of action that are relevant to policy research. He then states and discusses ten principles governing policy research. This is an insightful paper.

examination did not show a discarding of the earlier concepts as modifications emerged. Rather, there came to be multiple meanings of "equality of educational opportunity", different ones held by different segments of the population. What became clear was that, if this survey was to be responsible to the society and its representatives who directed that it be done, it must necessarily examine a number of different conceptions of equality of educational opportunity... the appropriate task of the policy researcher in formulating his research problem...is to carry out a sensitive and accurate translation of societal interests - potential or expressed - so that the very framing of the research problem will not myopically or intentionally favor interests of certain segments of society or further his own ideology. The research results might very well strengthen the hand of certain societal interests, as was certainly the case for Equality of Educational Opportunity; but the design must be such as to allow the possibility of results that benefit any interested parties - whether or not their interests coincide with those of his immediate client or his own values (Coleman, 1972, pp. 7-9).

An internal staff memorandum was written which set forth five definitions of inequality of educational opportunity and indicated how the survey would proceed. The final report presented evidence relevant to all five of the definitions. Three of the definitions were concerned with inequality of inputs into the educational system - community inputs (e.g., per-pupil expenditure, quality of teachers, libraries), racial and ethnic composition of the schools, and social and psychological characteristics of the students and their environment. The other two definitions were concerned with educational outputs (e.g., student achievement) from the educational system. These definitions led to an inputs-outputs design in which inputs and outputs were measured and the effect of inputs on outputs were related. The educational process and institutional structure which intervene were regarded in effect as a black box into which various inputs are placed and out of which various outcomes result.

Colemen has stated that the major virtue of the study was that it had an impact in shifting policy attention from its traditional focus on comparison of inputs (the traditional measures of school quality used by school administrators: per pupil expenditure, class size, teacher salaries, age of building and equipment, and so on) to a focus on outputs, and the effectiveness of inputs for bringing about changes in outputs. If the study had focused on input inequalities, its irrelevance would have gone unnoticed, as policy-makers busily worked to eradicate relevant as well as irrelevant inequalities (Coleman, 1972, pp.9-10).

Coleman has viewed input variables as being of two types, policy variables which are those amenable to policy control, and situational variables which can also affect outcomes but are not subject to policy control. Reports of the effects of situational variables give the client no basis for action. However, situational variables need to

be examined for they confound or distort the effect of policy variables on outcome variables. The distinction between the two types of input variables affects the questions posed by the research and the analyses carried out.

In the survey an example of policy variables was school facilities whereas a child's family background characteristics were situational variables. In regression analysis, which was used extensively in the report, these latter variables were used as statistical controls.

> The plan of the analysis was not appropriate for study of the relative effects of background and school variables, but it was correct for study of the relative effects of the different school variables. Since it was the latter that were the policy variables, while background variables were situational variables, this was as it should be (Coleman,1972, p.6).

The pathbreaking quality of the Coleman Report was the finding that, after removing (statistically controlling) the effects of student background variables on academic achievement, the remaining variation in achievement had little relation to variation in school facilities.

In order to examine and test the relationships between the variables for children of different races, religions, and social classes, some 570,000 school pupils and 60,000 teachers completed tests of verbal and mathematical ability and filled out questionnaires, reporting in some detail on their home environment and economic status. Information on the facilities in some 4,000 schools were gathered in elaborate detail.

The non-profit corporation, Educational Testing Service, was the contractor for the major school survey. It provided existing published tests for use in the survey and carried out the administration of these tests and of the special questionnaires developed by the staff of Mood's office. Educational Testing Service conducted the computer analysis in accordance with supplied specifications.

Reactions to the Administration of the Survey

Time - a property of the world of action relevant to policy research - placed pressure on the Office of Education. The project had been the responsibility of the lawyer Seeley for nearly a year before being placed in Mood's charge. Section 402 stated that the Commissioner of Education should make a report within two years to the President and Congress. The deadline was July, 1966.

In April of 1965, Commissioner Keppel arranged a meeting of a group of school superintendents[1] in order to ascertain their reactions to the survey testing since traditionally there had been opposition to federal testing programmes.

---

1) A school superintendent is administratively responsible for a group of schools situated in a local taxing district.

> Mood described the meeting as a catastrophe. The superintendents recited the litany of complaints about testing. It would be an administrative mess. Tests would lead to unsophisticated and invidious comparisons between school districts. Civil rights leaders in northern cities would subpoena the data for use in court suits then beginning to challenge patterns of de facto segregation in the North (Grant, p. 20).

Most of the reaction of the superintendents was due to the racial issues. Several participants at the meeting refused to take part in the October testing. Apparently Keppel became concerned that reaction to the projected testing might adversely affect the vote of congressmen in relations to the 1965 Elementary and Secondary Education Act which was nearing completion of its passage through Congress. This Act was to give massive funding and to break long-standing opposition to federal aid to elementary and secondary schools. Keppel temporarily placed the survey on ice.

A month later the Act was passed and Mood and Coleman were able to proceed. The massive organisation required for the administration of the thousands of tests and questionnaires across the country began. The co-operative support of hundreds of school officials at every level of education was necessary. Almost 20,000 school teachers were to carry out administrations in their classrooms.

Keppel wrote to state education officials on June 10 enlisting their co-operation; indicating that the aptitude and achievement testing part of the survey would be voluntary for local school systems; stressing that test results would not be used to compare achievement across systems or cities or states and that results would not be reported by individual schools or school districts; requesting that a state education official be designated to assist the Office of Education in developing suitable arrangements with local school systems; and making himself and Mood directly available by free phone calls for questions. In a similar manner, in a letter of July 13, Mood requested the co-operation and assistance of school superintendents. Mood indicated that in the questionnaires, questions about race, religion, and family background would be asked in order to illuminate issues raised in the Civil Rights Act. He further indicated that the questionnaire would make it clear that each student or teacher was fully free not to answer those questions which are felt to be objectionable. Both Keppel's and Mood's letters indicated the Congressional directives under which they were proceeding. Mood and Coleman, with full backing from Keppel, met with groups of superintendents in an attempt to answer their criticisms and reduce opposition.

In an interview several years later, however, Coleman felt that greater cooperation could have been obtained (more than a tenth of the target school districts refused to cooperate, including Chicago, Los Angeles, and the entire state of Florida). He blamed this on poor relations between the U.S. Office of Education and the states, a small liaison staff, and inadequate sensitivity to political problems. Perhaps greater effort to involve local superintendents in planning the survey and painstaking attempts to brief schoolmen through regional meetings may have paid off in terms of increased participation. (1) On the other hand, such a strategy could have increased grassroots opposition to the survey and might even have killed it altogether (Grant, p. 21).

(1) Footnote - "Again, perhaps the data gathering on race was a key factor. Project Talent, a more cumbersome study, received much better cooperation. Scholars are not unanimously in agreement, but the preponderant opinion is that the lack of participation of some school districts did not significantly affect the results".

The tests and questionnaires were administered on three days in October, 1965. As with the best of organized large-scale school surveys, there were the seemingly inevitable one or two horrible mixups, including a batch of forms filled out by school superintendents that had to be mailed out all over again because of errors in the forms. This occurrence did not increase superintendents' goodwill or confidence in the survey.

A controversy began to develop in the press about some of the attitudinal questions in the survey. Questions designed to elicit information about white children's racial attitudes were singled out by critics. Inquiries about background factors such as "who acts as your father?" also enraged opponents of the survey. A plaintive headline over a story in the Cincinnati Enquirer was not untypical of adverse press comments: "Must Johny Tell All?"

In Washington the fuss produced a number of inquiries from the press and congressmen that worried the survey staff. Since the data had already been collected, the controversy did not upset the survey. But a considerable amount of staff time was absorbed in reassuring critics of the confidentiality and anonymity of the data.... There was a feeling among many of those connected with the Coleman survey that they had just barely escaped the political rapids without upsetting their boat. It is quite likely that the experience also made Mood and Coleman more willing to listen to those who would later warn of the political dangers that might grow out of misinterpretation of their work (Grant, p. 21).

Eight months were left for the data on thousands of completed test and questionnaire forms to be organized and fed into computers; for complex analyses and comparisons to be carried out and interpreted; for the writing of the report.

The Findings - Their Release and Whitewash

The Coleman Report substantiated some commonly held beliefs but not others. It was the findings related to these latter beliefs that suffered the whitewash. As was expected, the Report showed that American school children were to a large extent separated by race. In the first few months following the survey administration and before the results of analyses began to emerge the correspondence

between Coleman's staff and the advisory committee appointed by Keppel shows that Coleman expected to find striking differences in the kinds of school attended by whites and Negroes and that these school characteristics would be highly related to differences in achievement between the two groups. This expectation was not substantiated. Three major political factors operating at the time made these latter findings unwelcome.

Firstly, the most dominant political pressure was concerned with the desegregation of the races in the schools. This often required the transportation of students by bus from one school to another. The lawyer, David Seeley, from whose hands the conduct of the survey had been taken, was in charge of the politically sensitive enforcement of the federal school desegregation guidelines. In addition, the Civil Rights Act of 1964 had empowered the federal government to cut off funds to segregated programmes. Mounting attacks from southern congressmen over the school guidelines and Mayor Richard Daley's[1] ire over Commissioner Keppel's attempt to cut off federal funds for Chicago because of segregation in the public schools there, contributed to Keppel's decision to step down as Commissioner in December, 1965. In January, 1966, Harold Howe II was appointed as successor. However, Howe issued even stronger desegregation guidelines. Keppel indicated to Howe that the survey could cause problems.

If it could be shown that school facilities and programmes of white and Negro schools were grossly unequal, then this could provide support for the clear legal mandate for desegregation. The survey found no major differences within geographical region between schools attended by Negroes and whites. In relation to desegregation, it was more important to achieve a mixture of students by socio-economic class and family background than by race since the former were significantly related to academic achievement. This did not provide support for the racial integration of poor blacks and poor whites in the rural South which was a principal target for integration at the time. Furthermore it was found that school factors did not markedly affect academic achievement. This did not advance the case for desegregation in the South or the North.

> In the North, it was argued that the promise of modern superschools with the latest equipment and modern curricula (and implied higher achievement for all) would be the sugar coating that would induce middle-class white parents to swallow the integration pill (Grant, p. 23).

---

1) Richard Daley, the Mayor of Chicago, is a very powerful force in American political life.

The second major political factor was the advent of the first federal aid programme which had been provided under the billion dollar Elementary and Secondary Education Act. The major part of the money was being spent on compensatory education programmes in the schools. The fear was that Congress might cut out the funds since Coleman's findings did not provide support for such programmes.

> Thus the impulse in the bureau managing this new program was to minimize the central finding of the Coleman Report, namely, that the schools in the ghetto, regardless of whether their programs are fat or lean by traditional measures, seem to have little influence on a child's achievement that is independent of his immediate social environment (Grant, p. 24).

The third factor was the concern that the Report could be used for racist arguments. White and Negro achievement scores were different but not their schools. The argument might be put forward that this showed that Negroes were intellectually inferior and that increasing school expenditure for them was a waste of money.

One might be led to wonder about the values of Coleman and what possible effect they might have had on the findings. Gerald Grant (1973, p.20) describes Coleman in the following manner.

> During the early days of the civil rights movement, Coleman and his family had picketed a segregated park in Baltimore, and he had sent his son to be among a handful of whites at a nearly all black YMCA (1) summer camp. But there is nothing of the idealogue in Coleman's personal style or his scholarship. He is a man with few flourishes in his manner or his work. He exemplifies Max Weber's ideal of the social scientist whose values properly determine the kinds of questions he asks, but whose scholarship commits him to look dispassionately at the results.

Coleman (1972, p. 14) has made the following statement on the role of the investigator's values in relation to policy research.

> ...acceptance of the research problem, an act that is still within the world of action, should be governed by the investigator's personal values, not disciplinary values of objectivity and truth. Both the willingness to accept the problem for study and the establishment of conditions of transmitting research results back into the world of action should be governed by these personal values. The execution of the research then must be governed by disciplinary values, which regard the investigator's own personal values and interests no more highly than those of other interested parties. Finally, the transmission back into the world of action must be conceived in two parts: first, reporting of research results, which must be done objectively and openly according to the disciplinary values; and second, expression and use of those research results in the world of action. In the latter part, the person who carried out the investigation may be an advocate, just as can any other person in the world of action. In the former, he cannot, any more than he can in the execution of the research. It may be difficult to separate these two capacities, but it is necessary to do so. For if it is not done, then the policy research loses its value for all interested parties.

---

1) This refers to a recreation camp run by the Young Men's Christian Association.

After the Report was released, the strong advocacy of Coleman was a factor which finally helped to make it an instrument of policy. During the writing of the Report, Coleman tried to resist the pressure of advocates within the HEW department who attempted to have the Report selectively present results to suit their interests. Only in relation to the Summary Report is it probably true to say that Coleman's attempt was not successful. The Summary Report occupies 32 pages at the beginning of the Report. It was the Summary Report which was released to the press in July, 1966, and not the full report, which was published the following month.

The Summary Report avoids raising some of the implications of the findings. Mosteller and Moynihan (1972, p. 29) state:

> The question of the "relation of achievement to school characteristics" is not raised until more than halfway through the summary, and then in terms that scarcely seize the attention or rouse the imagination of the unsuspecting reader.

The relevant paragraph of the Summary Report (Coleman et al., 1966, pp. 21-22) is as follows :

> The first finding is that the schools are remarkably similar in the way they relate to the achievement of their pupils when the socioeconomic background of the students is taken into account. It is known that socioeconomic factors bear a strong relation to academic achievement. When these factors are statistically controlled, however, it appears that differences between schools account for only a small fraction of differences in pupil achievement.

The way in which the highly misleading summary came to be released involved a range of concerns related to the three political factors previously discussed. During February and March of 1966, Coleman began to discern from the analyses that the characteristics of the school did not seem to make much difference in the education of the child. This became known in the offices of the Department of Health, Education and Welfare (H.E.W.) and caused concern, especially to those connected with civil rights responsibilities. The July 1 deadline was approaching, and Coleman began to feel this pressure in April. Both Coleman and Mood were surprised at what was coming out of the analyses. Coleman in his memos to Mood urged that an extension of time be given.

> However, the initial reaction of Mood and Coleman may have given their bureaucratic adversaries the feeling that the researchers were unsure of their conclusions. The upshot of the request for a postponement was a decision to publish the summary document by the deadline established by Congress and to publish the full report....later. With that key decision made, the bureaucratic conflict began to intensify (Grant, p. 25).

On June 2, some rough drafts of a few chapters became available within the Department. The civil rights activists were highly dissatisfied since what came through as the main conclusion was that school caracteristics had little to do with pupil achievement. According to Seeley, Coleman was "taking bits and scraps of evidence

and forming them into what will surely be read as conclusive proof of this astonishing fact" (U.S. Office of Education archives). Seeley, one of the most dissatisfied members of the civil rights group, made it clear to Mood and the Office of Education executive that the report should highlight the status of segregation in various parts of the country.

Jeanette Hopkins, an editor for the publishing house of Harper and Row, was asked to prepare the summary report. The resulting summary showed a writer's flair for the dramatic, little understanding for the methodology of the survey and relied heavily on a few small-scale studies that Seeley had originally initiated for the survey. In Mood and Coleman's view, her summary was a disaster and they succeeded in having it quashed.

Mood, who had been working quietly on his own summary, dropped everything else to work on it. The resulting summary, while giving a cautious interpretation of the major findings of the survey, gave its major emphasis to the relation of school characteristics to achievement. The distribution of Mood's summary report brought strong reaction from a number of officers of H.E.W. who thought it would provide a major refuge for those who opposed civil rights activity, would bring "great distress" on the Office of Education, and could generate a press catastrophe.

> Coleman at this time had barricaded himself in a motel in southwest Washington, a few blocks away from the Office of Education, where he was working round the clock on the key chapters of the full report. Meals were sent in as he grappled with new analyses sent down each day by Greyhound bus from the computer center at ETS in Princeton. Mood was coordinating other aspects of the study (Grant, p. 28).

Howe was persuaded to request Helen Rowan, editor of the Carnegie Quarterly,(1) to try to write a more acceptable version of the summary. This was written under extreme time pressure, with consultations between Mood and Coleman. This third version was the one that was released to the press and became the beginning of the full report. While Rowan's version did not overtly mis-state anything, it did not place any emphasis on the relation of school characteristics to achievement as had been the case with Mood's version.

The press conference introducing the report was held on July 1 at 1:30 p.m. It was unusual in two respects. The Summary Report was not made available to the press until just before the press conference began. It is usual for the government to give reporters a few days to absorb the contents of complex technical reports. Secondly, there was no press release from the Office of Education highlighting the import of the survey.

---

1) This is published by the Carnegie Corporation, a philanthropic foundation. Each issue describes some of the Carnegie-supported projects, primarily in the area of education.

Howe conducted the press conference. He read excerpts from each section of the Summary Report, giving emphasis to none and refused to entertain any questions about policy implications. Howe remembers not fully understanding the Summary at the time he presented it.

> The question period following Howe's presentation revealed that many newsmen were at least equally baffled. The summary seemed chock full of tables but with very little news. Most appeared hesitant to draw any conclusions from the data when an "expert" like Howe wouldn't.
>
> Although there really was not much news in the fact that American schools were highly segregated, it seemed to a number of reporters to be the most prudent news peg. Certainly it was safer than drawing the wrong inferences from the rest of the data (Grant, pp. 29-30).

The report by the Associated Press news service, which appeared in the hundreds of American newspapers that depend on the wire service, was typical:

> Washington (AP) - The Office of Education has told President Johnson and Congress that a nationwide survey of education shows American children are attending public schools which are largely segregated.
>
> "We have now documented what we always believed", Education Commissioner Harold Howe II said. "American children attend schools where almost all of their fellow students are of the same racial background as they are" (Tom Seppy, Biloxi Herald, July 2, 1966).

In August, the full report was published. It was 737 pages of technical objectivity, including graphs, tables and statistical analysis. It was accompanied by a second volume of 568 pages of means, standard deviations, and correlation coefficients. Considering the amount of time that was available for this immense survey, the final report was a monument to physical and intellectual exertion. However, the report is a difficult and time consuming document to read and grasp. To understand the nature of the analyses requires some degree of statistical sophistication. Much of the report is to some degree inaccessible to non-specialists. Policy implications are not spelled out and no remedies are proposed. It is a document for the social science disciplines.

After the press conference, Coleman was interviewed by writers from the Washington Post and the Chicago Tribune and they picked up and reported the "hidden" significance of the study. However, these reports went unnoticed by the scholarly community which relies heavily on The New York Times which had printed a story conforming closely to the Summary. The educational establishment on the whole, chose to overlook the Coleman Report, perhaps because it did not conform to their assumptions about the effects of school facilities, curriculum reform and teacher training on school achievement.

A memo to the Commissioner, written by the Director of Public Information at the Office of Education, shows great pleasure over the press coup:

> What can be said from a public information point of view is that in the months since its issuance the report has created no public relations problems. On the credit side, it has provided you with some telling points for your speeches and has reinforced the Office of Education to accelerate school desegregation (U.S. Office of Education files).

In August, with the publication of the full report, Howe asked for reaction from head personnel within his Office. Most responses were fairly neutral, recommending minor adjustments or expansion of existing programmes.

In October, Howe called a group of scholars and other educators together to advise him about what the Office of Education ought to be doing about the Coleman Report. Congress, which was about to pass an extension of the Elementary and Secondary Education Act, was not affected by Coleman's findings. No implications of the Report had been passed on by the Office of Education. Key staff advisors to congressmen had not read the Report, it being a weighty document.

> Howe termed his meeting with the academics frustrating. "All I got was a lot of muddy water", he said.(1) Two years after it has been completed, the Coleman Report had no discernible effect on any program within the Office of Education (although it was cited occasionally by Howe in speeches), and plans for a follow up study were quietly dropped when no one pressed for it. Mood left the agency. Coleman returned to the academy. The Johnson White House remained cool to Coleman's findings" (Grant, p. 32).

Writing in 1972, Coleman (p. 12) states:

> The values governing transmission of research results back into the world of action, and thus the conditions of that transmission, are determined by conditions of acceptance of the policy research problem. The transmission may be largely controlled by the investigator of the discipline, and thus governed by disciplinary values, or largely controlled by the client who formulates the research problem, and thus governed by the values of an interested party to the action. The principal point at issue will ordinarily be the open publication and dissemination of research results, following disciplinary values, or constraints upon dissemination, following the values of the client. There can be no general statement that it should always be controlled by one or the other party, no general statement that it should always be open dissemination. The discipline may attempt to argue such a principle in a "code of ethics", but that is merely the restatement of the disciplinary values. Whether these should override the client's interest in controlling dissemination of research results is a question that must be answered afresh for each case, through negotiation between client and investigator. The outcome will depend, at least in part, on the degree to which one can justify his values to the other.

The Effect of the Coleman Report on Policy Makers

It was nearly four years from its publication that the Coleman Report began to exert a significant effect on federal policy. Daniel P. Moynihan played a strong hand in this, second only to that

---

1) Interview with Harold Howe II.

of Coleman. In January 1969, Moynihan moved to the White House as the key adviser to President Nixon on urban affairs. He was the principal author of President Nixon's Message on Education Reform of March 3, 1970.

> The message was grounded on research evidence, as no previous education message from the White House had been. It was the Coleman Report writ large. And in terms of officially stated policy - not, of course, the same thing as an actual change of existing law and practice - it was a dramatic reversal of the whole thrust of the federal educational programs of the Kennedy-Johnson years (Grant, p. 33).

The message indicated that young people may be learning more outside school than they learn in school, that compensatory education programmes were ineffective, and that a direct connection between school characteristics and student achievement did not exist.

There was a concern, however, that the line of argument in the message could be used to justify reduced financing of education. This in fact did occur in relation to the total budget for the Office of Education which was reduced by more than $ 450 million from the previous year. In relation to federal aid to schools this was not possible, however.

> How much of Nixon's embrace of the Coleman Report was political expediency and how much of it was the result of Moynihan's advocacy of what he believed was a powerful piece of evidence that could not be discounted? The answer cannot be given with much certitude. But assuming it was all expediency, what would be the point of developing the argument in such detail? Why risk alienating the powerful educational bureaucracies who do not like being told their rationales for huge federal aid to education programs are a house of cards?
>
> The irony of the President's message was that in the area where he leaned most heavily on the Coleman Report, he was least successful in achieving his goals. The Johnsonian (1) education legislation had become bread and butter school programs for every congressional district in the country, and Congress seldom shuts down popular bakeries. A month later The New York Times reported: "With considerable reluctance, President Nixon signed today the most expensive education legislation ever passed by Congress - a bill authorizing $ 26.4 billion over three years for elementary and secondary education".(2)
>
> Major shifts in policy cannot realistically be expected to come to fruition immediately...Coleman made several appearances on Capitol Hill that spring, and he appeared in the news so often that The New York Times ran a "Man in the News" profile about him...Moynihan played a key interventionist role as an advocate for the Report - not only in the White House but also by supporting reanalysis of the Report at the U.S. Office of Education

---

1) A reference to President Johnson and to the Elementary and Secondary Education Act of 1965.

2) Robert B. Semple, Jr., "Nixon Signs Education Bill with Reluctance," The New York Times, April 14, 1971.

and by citing the Report in meetings throughout the executive and legislative branches of government. The kind of visibility Coleman and his research had achieved in Washington, D.C. gave the Report more weight elsewhere (Grant, pp. 36-38).

President Nixon's major education message on March 3, 1970, made only passing reference to desegregation. The next day, John D. Ehrlichman, then Assistant to the President for Domestic Affairs, indicated that changing the racial makeup of the schools for a purely social end was a misuse of the school. He stated that the President saw the schools as an instrument of education primarily. Ehrlichman was testing the reaction to future possible policy.

Four days later in a front page interview in The New York Times, Coleman made a response to Ehrlichman without specifically naming him.

> School integration is vital not merely for some vague generalized social purposes, but because it is the most consistent mechanism for improving the quality of education of disadvantaged children (Rosenthal, 1970, p. 1).

Coleman went on to say that there were other ways to make integration work than by busing students from one school to another. In the course of the interview, Coleman summarized the major implications of the Coleman Report with regard to the integration issue, pointing out that integration can and does have major effects so long as middle-class children remain in the majority in a given school. Coleman said that he hoped that when the President issues his desegregation statement, "he will show federal commitment to investing some resources to insure that integration occurs and to making it work" (Rosenthal, 1970, p. 26).

A week later as a result of a Presidential request, Coleman became a consultant to a new Cabinet Committee on Desegregation. On March 19, Coleman spent an hour talking directly with the President. While Coleman had an important role, he did not write the President's policy statement on school desegregation which appeared on March 24, 1970. Coleman's influence was evident in two ways. Firstly, the message contained the first clear Nixon statement of the educational benefit of integration and secondly, other ways of achieving integration besides busing were discussed. However, the message expressed a preference for spending on compensatory programmes, and facilities, and materials. The President had preached nearly the exact contrary of what he had said in his education message three weeks earlier.

Coleman continued to state his views which were reiterated and amplified in Congress a month later in the hearings of the Subcommittee on Equal Educational Opportunity. Coleman's message was increasingly being heard in the White House, specifically in the Cabinet Committee on Desegregation of which Vice President Spiro Agnew was chairman. Two months after the President's message on school

desegregation, the Emergency School Aid Act was announced by the White House. The Act bore the imprint of Coleman's advocacy. Coleman felt that the bill had moved much further toward a policy that would create strong and stable school integration than had been clear in the President's March 24 statement on desegregation policy.

Coleman played a significant role in drawing up the administrative guidelines which would govern how the monies of the Act could be used. He appeared before the House Committee on Education with regard to hearings on the Act.

> ...Coleman was treated by the committee (and in many ways he had become) as the principal administration spokesman for a bill which he had an important role in shaping. It was an unusual and unusually effective presentation for a social scientist whose Report had been virtually ignored by the House Committee four years earlier (Grant, p. 50).

Coleman, in 1972, made the following statement about the use of the Report by policy makers:

> The experience of this research has suggested several things to me concerning the relation of research to social policy. One is that a major reason for non-use of research results in policy lies in the fact that it's not commissioned by a policy-making body. A different pattern existed with Equality of Educational Opportunity. It was not initiated by a grant from the OE Bureau of Research, but in implementation of legislation from the congress. That fact in itself created the presumption of policy relevance, and automatically created audiences among policy-making bodies, from the presidency to local school boards.
>
> The actual use of the report has been more at the local than the state or federal level (Educational Researcher, 1972, pp.13-14).

The Effect of the Coleman Report on the Courts

On May 17, 1954, in the decision known as the Brown v. Board of Education of Topeka, Kansas, amidst an atmosphere charged with emotion, the Supreme Court ruled that state-imposed racial segregation in the schools was no longer legal. Almost all the voluminous arguments presented on behalf of the plantiffs were designed to show that racial segregation in education is, itself, a form of inequality. There was angry reaction to the decision. Segregationists and states' rightists formed one alliance, while civil rights advocates and liberals formed another. Following this decision, a number of civil rights cases were brought before the courts. In the late 1960's and after, the findings of the Coleman Report were introduced as evidence or expert testimony in a number of cases. The Report in fact exerted an effect on the courts before it had significant effect on federal policy.

Coleman, in an interview in 1972, stated :

> The study has had impacts I never would have predicted. The primary impact of the study as far as equality of educational opportunity is concerned has been in the courts.
>
> Issues such as school segregation, which the report didn't explicitly address, have left judges looking around for some kind of evidence on which to base a ruling about what constitutes equal educational opportunity (Educational Researcher, 1972, p. 13).

In an early case in which this Report was used, Hobson v. Hansen, in the District of Columbia, Coleman testified in court. The judge cited Coleman's findings in coming to his decision to abolish tracking, integrate faculties, and to increase integration among pupils insofar as it was possible in a city in which the public schools were nearly 90 per cent Negro.

In the case, Beckett v. School Board of the City of Norfolk, both the plaintiffs and the defendants used evidence from the Coleman Report in their arguments. The defendants argued that the Coleman data showed all children do better academically in schools that are predominantly middle-class in enrolment. Here the optimal desegregation plan in Norfolk would mean that each school in the system would achieve an enrolment of middle-class children proportional to the total number of such children in the system.

> The plaintiffs argued that while there may be some benefit from a social class mix in school, the data showed the greatest effects could be attributed not to one's classmates but to the social milieu of the family and home. In reply, the court rejected this line of argument because even if it was true, it was "hardly susceptible of control by any school system". That is, the schools could not assign children to different families (Grant, p. 52).

The extent to which Coleman's work had an effect is indicated by the following statement of the court:

> We cannot believe that the Supreme Court, in requiring "desegregation" has merely ordered a mixing of racial bodies without consideration of the social class factor.
>
> What the courts are now overlooking is that the primary variable is social class, not race.
>
> In effect...(the witness) concludes that unless an intelligent approach is made to the problem of desegregation from a social class standpoint - and not solely through the mixing of racial bodies - desegregation will be a complete failure . (1)

In a 1970 Denver case, Keyes v. School District Number One, the court, largely following Coleman's findings, saw no effect on educational opportunity traceable to age of building or size of site, although it recognized that "such factors may aggravate the aura of inferiority which surrounds the school". (2) The court asked Coleman

---

1) Beckett v. School Board of the City of Norfolk, 308 F. Supp. 1274, 1969.

2) Keyes v. School District Number One, Denver, Colorado, 313 F. Supp. 61, 1970.

to testify as to what remedy would provide the most nearly equal educational opportunity and ruled that the only constitutionally feasible and acceptable programme, the only one approaching "substantial equality", is a system of desegregation and integration which provides compensatory education in an integrated environment.

Coleman has made the following statement about the use of social science evidence by the courts:

> It's probably not appropriate to say on achievement grounds alone that segregated schooling does not provide equality of educational opportunity. There is not sufficient evidence to show that the kind of benefits to lower-class children that arise from a socio-economically heterogeneous or racially heterogeneous school can't also be provided by other means. I don't think a judge can say there is prima facie evidence of inequality in educational opportunity on achievement grounds if there is school segregation. In this sense, I think judges have looked at that study and used the results more strongly than the results warrant.
>
> I remain uncertain about the appropriate role of social science evidence or statistical evidence in relation to the courts. The concept of evidence by lawyers or judges is very different from the concept of evidence in social science. When results show that certain kinds of attendance patterns provide higher achievement for children from lower socio-economic levels, as our results did, the results ought to contribute to the question of whether schools should be integrated, and to the decision of how much effort should be put into school integration. But I don't think that a judicial decision on whether certain school systems are obeying or disobeying the constitution ought to be based on that evidence.
>
> At the same time I think the report has been underutilized by legislatures and school boards. The results of the report are appropriate for legislators and school boards in encouraging the kind of student body mix which can provide achievement benefits (Educational Researcher, 1972, p. 13-14).

The Effect of the Coleman Report on Academia

The Coleman Report was the first major influential study of the effect of schooling. The most startling finding to the academic world was that school factors did not seem to have much effect on student achievement whereas student background factors did. Probably because of the academic's commitment to schooling and education, this finding presented a challenge to the social science disciplines to determine what factors are important in influencing the outcomes of schooling. This challenge still exists today.

The Coleman survey set the major precedents for school effects analyses as well as being the prime exemplar for subsequent criticism. Although it caused the greatest amount of controversy, the analysis of the relations between school characteristics and achievement occupies only 40 pages out of a total of 737.

The Report created widespread interest, discussion and writing in the academic world. The Report had relevance to a whole range of university departments - sociology, education, mathematics, law,

political science, economics, philosophy, psychology. Reactions to the Coleman Report can be found in writings in each one of these areas.

The analyses of the data were done under great pressure of time. However, in the Report itself, care was taken to qualify principal findings, to point to alternative explanations of the data, and to disclaim certainty on many of the results. Inevitably, because of its magnitude, nature and implications, the survey came under critical attack. Inevitably, there were disagreements over the merits and inadequacies of the survey.

> Professional judgments were made in the survey design and analysis that are open to dispute. This is not always a matter of being right or wrong, but simply of the absence of professional consensus on the points involved. Inevitably a fair amount of technical controversy arose over the study. Not a great deal was published, but enough to serve as ground for suspicion for any so disposed. A range of criticisms was made. The sample was not properly done. The nonresponses were too many. The number of school systems refusing to cooperate invalidated any results. The reliance on school administrators to report accurately the facilities in their schools was naïve. Etc., etc. The subsequent analysis also came under scrutiny, perhaps most notably by Samuel Bowles and Henry M. Levin... Bowles and Levin reached conclusions somewhat like that of the Scottish verdict "Not Proven".... In the end Bowles and Levin came near to suggesting that the whole thing needed doing over again (Mosteller and Moynihan, 1972,p.33).

Bowles and Levin (1968, p. 23) state:

> ...some of the Report's most widely publicized findings concerning the determinants of scholastic achievement, namely, those relating to the ineffectiveness of school resources, the influence of student peers, and the effects of integration, are not substantiated by the evidence. We have attempted to show that both the measurement of the school resources and the control of social background of the student were inadequate, and that the statistical techniques were inappropriate. By no means do we wish to suggest that the actual relations are the opposite of what the Report concludes or that further research will not substantiate some of the Report's findings; but until better evidence is found, we will have to remain agnostic about which relationships prevail.

The outcomes of education are complex, multifaceted phenomena. To measure the significant outcomes of education is a virtual impossibility. Those which are most capable of measurement are usually studied. Consequently, educational outcomes are almost exclusively measured by standardised tests of academic achievement, as was the case of the Coleman Report. While no one would deny the importance of affective and social outcomes they are extremely difficult to assess. Furthermore, standardised tests only measure a few facets of a student's total academic achievement. Unfortunately, policy makers and the courts are often not aware of this inherent inadequacy of much research relating to educational outcomes. Such research may provide rough guidelines but rarely certainties. Unlike Bowles and Levin, many commentators on the Coleman Report, while pointing out its inadequacies, would usually indicate its policy implications

as if these inadequacies did not exist. Without denying its path-breaking qualities, it is probably true to say that more reliance has been placed on the findings of the Coleman Report than is justifiable.

In the preface to the Report Mood indicated that only brief analyses had been made by the Office of Education in the few months available since the data were collected. He stated that the Office of Education would make all the data gathered by the survey available to research workers. This was an original step by the Office of Education which made great efforts to assist scholars in their reanalyses, whether the researchers concerned were friendly or hostile to the findings of the original report.

In 1966 Moynihan and Thomas Pettigrew initiated at Harvard University an inter-disciplinary seminar which led to a major reanalysis and eventual scholarly buttressing of the Report. There was much involvement of the academic staff at Harvard.

> What had been expected to be rather a specialized undertaking soon attracted widespread interest... Here was a subject of such inherent intellectual and social interest as to challenge the best of minds. Before long the evening seminar assumed near conference proportions, with fifty to sixty faculty members, and other interested persons, swarming about among panels, committees, groups, and subgroups. It was a spirited and rewarding experience (Mosteller and Moynihan, 1972, p.ix).

The results of these efforts appeared in 1972 in the volume On Equality of Educational Opportunity, edited by Frederick Mosteller and Daniel P. Moynihan. James W. Guthrie and Thomas C. Thomas (1973, p. 602, 605) make the following comments on the book:

> This book is principally a technical reanalysis of the quantitative findings of the Coleman Report. A number of revisions are made, but they are consistently in magnitude rather than direction;...The Coleman Report has been criticized for the absence of an adequate conceptual model, for weaknesses in its sampling procedures, for failing to pose valid survey questions, and for employing inappropriate analytical procedures.(1) Because the Harvard analysts utilized the same data, they too have fallen prey to many of the same criticisms.(2) Moreover, because much of the reexamination was conducted independently by individual

---

1) See, for example, Samuel Bowles and Henry M. Levin, "The Determinants of Scholastic Achievement: An Appraisal of Some Recent Evidence", Journal of Human Resources, Winter 1968, pp. 3-24.

2) See, for example, David Wylie's review of On Equality of Educational Opportunity in Science, September 1972; Gerald Grant's essay review, "On Equality of Educational Opportunity", Harvard Educational Review, February 1972; and James W. Guthrie's "What the Coleman Reanalysis Didn't Tell Us", Saturday Review/Education, July 22, 1972.

> researchers, the various component chapters of the complete volume do not always agree with each other.(1)...The reanalyses tend to confirm the basic findings of the initial report....A popular impression of the Harvard reanalyses is that they demonstrate that additional resources for schools would be a waste of money. This impression is not supported by a careful reading of the book, nor could it be, since neither the Coleman Report nor the reanalyses undertook a comparative study of the uses of public resources, which would be required to support such a conclusion.... Perhaps the most important implication of the book, however, is to illustrate the pressing need, not for continued reanalysis of old and flawed data, but for incisive new policy-orientated data relating to a different kind of variable. Data are needed which tie together students, teachers, curricula, and resource levels.

A number of other reanalyses of the Coleman data have taken place. A team of analysts in the U.S. Office of Education headed by George Mayeske have produced two volumes, A Study of Our Nation's Schools in 1972, and A Study of the Achievement of Our Nation's Students in 1973. The purpose of the first study was to discover what characteristics of American schools are most closely related to school outcomes. This involved devising methods by which school influences could be distinguished from social background influence. The purpose of the second study was to find out which aspects of a student's background, whether alone or in combination with the kind of school attended, play the largest role in achievement. Another highly significant study which was an outgrowth of the Coleman Report, was that by Christopher Jencks et al. (1972), Inequality, A Reassessment of the Effect of Family and Schooling in America. One aspect of this study attempted to look at the long term effects of schooling and family background.

> Equality of Educational Opportunity addressed itself to some of the most difficult questions that our society faces; what are the determinants of different educational outcomes, and what is the relative importance of each of the relevant influences? Unfortunately, the survey that led to the Report was handicapped by a severe time constraint. It was also hampered by a more serious impediment, for the learning processes by which different influences alter achievement are largely unknown, and no set of data and statistical analyses can easily compensate for a missing theoretical framework.
>
> The Report has a distinct contribution to make, but not directly in the arena of educational or social policy. Rather, its strength lies in the fact that it has stimulated a great deal of thought and new research efforts to uncover the largely unknown and complex relationships among family, school, and community influences on one hand, and educational outcomes on the other. Further, it has provided some of the necessary data to test the new hypotheses

---

1) For example, Christopher Jencks in Chapter Two concludes that expenditures have little impact upon increments of student achievement (pp. 90-92). Conversely, David Armor concludes that raising per-pupil expenditure levels might benefit black students (p. 221).

that it has stimulated. In short, while the Report did not provide the answers, it has brought us closer to being able to use large scale research efforts as a basis for making intelligent policy decisions for our schools (Bowles and Levin, 1968, pp.23-24).

Beginning with the Coleman Report, large-scale statistical studies have failed to show consistent and important relationships between what goes on in schools and variations in student learning, as measured by cognitive achievement tests (Averch et al., 1972). These reports have upset the educational profession and have comforted and supported those who would reduce the resources allocated to education. Some see the reports of the lack of effect of schooling as having created a new mythology. They countermand common sense and seem to imply that belief in effective education is a delusion. Thus the search for school effects continues. The Coleman Report opened a Pandora's Box.

REFERENCES/SOURCES

AVERCH, H.A. et al. How effective is schooling?: A critical review and synthesis of research findings. Santa Monica, California: The Rand Corporation, 1972.

BOWLES, S. and Levin, H.M. The determinants of scholastic achievement - an appraisal of some recent evidence. The Journal of Human Resources, Winter 1968, III(1), 3-24.

COLEMAN, J.S. Telephone interview, November 18, 1974.

COLEMAN, J.S. et al. Equality of Educational Opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

COLEMAN, J.S. Policy Research in the Social Sciences. Morristown, New Jersey: General Learning Corporation, 1972.

Educational Researcher. Coleman on the Coleman Report - an interview. Educational Researcher, March 1972, 13-14.

GRANT, G. Shaping social policy: the politics of the Coleman Report. Teachers College Record; 1973, 75(1), 17-56.

GUTHRIE, J.W. and Thomas, T.C. I. Policy implications of the Coleman Report reanalyses. Phi Delta Kappan, May 1973, 602-605.

JENCKS, C. et al. Inequality. New York: Basic Books, Inc., 1972.

MAYESKE, G.W. et al. A study of the Achievement of our Nation's Schools. Washington, D.C.: U.S. Government Printing Office, 1972.

MAYESKE, G.W. et al. A Study of the Achievement of our Nation's Students.

MOSTELLER, F. and Moynihan, D.P. (Eds.) On Equality of Educational Opportunity. New York: Random House, Inc., 1972.

ROSENTHAL, J. School expert calls integration vital aid. The New York Times, March 9, 1970.

4. THE MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM

In the middle of 1968 a number of researchers joined the State of Michigan's Department of Education to staff a new office, the Bureau of Research. They were impressed by the "lack of reliable information about the level, distribution, and progress of education in the state" (Murphy and Cohen, 1974, p. 56). At the same time, they were aware that J. Alan Thomas (1968) had surveyed the state's school finance arrangements in 1967. Thomas had concluded, "There is great variation in the educational opportunities available to children in the state of Michigan" (as quoted in Kearney et al., 1970, p. 2). Those in the Bureau of Research thought that better information would be needed in efforts to equalize educational opportunities. They expected that an assessment of inputs and student achievement outcomes would show that school expenditures were highly correlated with student achievement. With such data in hand they could argue that aid for poorer districts would raise achievement for low-scoring students.

The staff of the Bureau was troubled by the fact that although data existed on numbers of schools, buildings, teachers, and pupils, and on monies expended, there was a lack of sound and adequate information on educational results. Without dependable information on educational results, decisions about what course Michigan education should follow had to be based on impressions at best. Therefore, they hoped to collect information which would serve as a basis for rational decision-making. This was one of the basic reasons given by Ralph Tyler who had independently been leading the drive for a National Assessment of Educational Progress. The writings of Ralph Tyler had influenced the staff of the Bureau.

By late 1968, the staff of the Bureau had already begun informal discussions about the lack of reliable information on the progress of education in Michigan. The discussions led to the preparation of a paper which outlined the problem and suggested that the Department undertake a state-wide assessment. They circulated the paper within their office and to Ira Polley, the State Superintendent of Public Instruction (the top educational officer in the state). Polley reacted positively. About two years earlier Polley and an Associate Superintendent had wrestled with the problem of equitable distribution of money provided by Section 4 of the State School Aid Act. Polley then began discussing strategies for implementing a state-wide assessment of student achievement.

In January, 1969, the State Board of Education - under what seemed to be growing pressure from taxpayers to justify increasing costs of schooling - moved decisively toward improving and regulating the quality of education in Michigan. The Board proposed an increase in state involvement in school accreditation, a function then performed by the University of Michigan. In his capacity as Superintendent of Public Instruction, Ira Polley also served as Chairman of the State Board. He opposed the idea. He suggested that a greater contribution to upgrading the quality of education in Michigan would be to develop tools for accumulating information which could be used as a basis for decision-making. Polley offered to carry out the assessment that his Bureau of Research had outlined. The Board accepted. But it did not approve the Bureau's request for three years to plan a comprehensive assessment programme. The Board insisted on more immediate results.

A further attempt was made by Superintendent Polley to obtain a three-year planning and development period. It then became apparent that one of the Board members was prepared to introduce a legislative resolution calling immediately for a state testing programme with achievement testing mandated in each school district for all children in grades 3, 6, 9, and 12. Standardized tests were to be used in the areas of reading, mathematics, and written and oral language skills. All test results were to be made public. The University of Michigan was already operating a voluntary statewide testing programme. Superintendent Polley believed that the programme proposed in the resolution would not provide information that would indicate where there should be increased state funding.

A compromise was reached. In March, 1969, an amendment was made to the three-year plan and an administration of "basic skills" tests to pupils at two grade levels was scheduled for the 1969-70 school year. Another portion of the plan allowed for the collection of background and input data. What the Superintendent's office particularly needed was socio-economic information to correlate with the group achievement test scores.

In the first public statement of purposes and procedures for the assessment (Michigan Department of Education, 1969) the following assumption was stated:

> ...the most important education-related problem facing the State - and indeed the nation - is the inequitable distribution of school district performance levels and their correlates (p. 48).

The ultimate goal of the assessment effort was:

> ...to provide members of the State Board of Education and the Legislature with information needed for allocating the state's educational resources to equalize and improve the educational opportunities for all children (p.49).

Looking further into the future a second goal, closely related to the first read:

> to improve the basis for educational decision-making over time (Kearney et al., 1970, p.2).

The Board passed a resolution directing the Superintendent to seek legislation that would provide a mandate and the necessary funds to undertake (a) a long-range planning and development effort, and (b) an immediate assessment of certain basic skills at one or more grade levels during the 1969-70 school year. After lengthy discussion and consistent urging by the State Superintendent, a line item was added to the Department of Education's budget bill that year. On August 12, 1969, Governor Milliken signed the budget bill with the assessment provision intact. The Michigan Educational Assessment Program then became law.

To sum up, the original idea for a statewide assessment came from the staff of the newly created Bureau of Research. Superintendent Polley subsequently proposed the assessment as an alternative to the School Board's proposal to increase state involvement in school accreditation. The School Board then passed a resolution directing the Superintendent to seek legislation providing a mandate and funds necessary to undertake the assessment. This was the original authorization. A line item was then added to the Department of Education's budget for the 1969-70 school year. The budget bill eventually passed both state legislative bodies in Michigan. The Governor ultimately signed the bill mandating the assessment. Thus an evaluation procedure was created by law to find out how well students from the state's school districts were achieving in different social and educational circumstances.

After the bill allocating funds for the Assessment Program was passed, only five months remained before the first tests were to be administered. One Bureau member has stated that the Bureau had less optimum time to crystallize the aims of the Program and construct instruments for assessment. A nationally recognized test publisher, the Educational Testing Service of Princeton, New Jersey, contracted to help set up the Program and develop the tests. With so rushed a schedule and some ambiguity as to purpose, the Bureau became greatly dependent on a test publisher over 500 miles away. The Bureau found itself in only partial control. In a casual moment C. Philip Kearney, Associate Superintendent for Research and one of the original members of the Bureau of Research, remarked that the Bureau had perhaps only 30 per cent control in shaping the Program that first year.

Before the first assessment was administered Ira Polley left the State Superintendency to assume a position at Michigan State University. John Porter, formerly an Associate Superintendent for Polley and a supporter of the assessment effort, assumed the office of Superintendent of Public Instruction in October, 1969.

The first assessment of basic skills was ready for administration in January, 1970. But before it was implemented statewide, there was an unexpected reaction to Section 1 of the Basic Skills Battery. Section 1, titled General Information, contained a number of questions designed to provide group measures of socio-economic status, pupil attitudes, and pupil aspirations. Civil rights leaders and parents of underachieving students - those whose children the Assessment Program was specifically designed to benefit - complained vehemently about "racism" and "invasion of privacy". The furor started in Detroit and spread to Lansing, the state capital. Numerous school officials assumed that they had the discretionary power to delete Section 1, and did. Reaction to Section 1 led to introduction of a resolution in the Michigan Senate calling for a moratorium on the Program.

An even greater disturbance was caused by several reading comprehension items. The items were based on two passages: the first, an excerpt from the F. Scott Fitzgerald short story "The Rich Boy"; and the second, a description of the human problems arising from industrialization, taken from a twenty-year-old social studies textbook. Conservatives saw these passages as "Communist propaganda, anti-free enterprise, and anti-American". Reaction to the two passages led to introduction of a resolution in the Michigan House of Representatives calling for a moratorium on the Program.

Spokesmen for the State Department talked with legislators in both houses and provided the assurances necessary to set aside both resolutions before they came to a vote.

Having administered the tests and having collected the data, the state officials were uncertain as to what to do with the results. Much to the disappointment of agency personnel, the data did not show strong relationships between school achievement and school resources. They had found no clear and persuasive evidence for reallocating school aid to the poor. The correlations were not made public. Reports which graphically illustrated the relative standing of the districts were published.

The Bureau was also faced with the difficulty of publicizing the achievement results without causing another political controversy. The initial design of the 1969-70 Assessment Program had established a three-part policy on reporting assessment results. First, the Assessment Program was designed to gather information not only on

performance measures but also on input factors. This meant that outputs would always be considered in relation to inputs. Second, the school districts of the state were categorized into five community types and four geographic regions. Therefore, no single school district would have to be identified in order to make comparisons among community types and/or geographic regions. Third, the Department stated publicly that it would neither prepare nor publish any ranking of Michigan school districts based on 1969-70 assessment results. "In effect, the Department had gone on record as stating it would not release data on individual local school districts to anyone but local school administrators - unless forced to do so by external factors" (Kearney and Huyser, p. 50).

Late in August, 1970, a report which included the local district's own assessment scores was sent to each of Michigan's public school districts. The State Superintendent encouraged each district superintendent to make the results public. Some superintendents did make the results public, others did not. In response to demands from legislators and irate citizens, eventually the results were made public to all.

The decision to release the results of the 1969-70 assessment to the public was reinforced by a political development. In the 1970-71 State School Aid Act, the Legislature allocated $ 17.5 million for the continuation of a school aid programme for those suffering a high degree of economic and cultural deprivation. Section 3 of the 1970-71 State School Aid Act required that to be eligible for funds, individual schools had to have a high percentage of poor and underachieving students. Criteria were to be based on the results of the 1969-70 assessment in Grade 4 of each school. The legislation, in effect, required that the State Department prepare a ranking of all schools in the state and be prepared to release the point scores of all schools falling in the bottom quarter on two of the assessment measures: composite achievement and relative socio-economic level. The Department released individual districts' test scores in October. The Department cites the Section 3 legislation as justification for releasing assessment data.

Not surprisingly, the Department's publication of results brought repercussions. Some legislators were happy because the results could be used to bolster their positions. Some were greatly unhappy. Many individuals were plainly disappointed because the results did not provide answers to key questions. Some school administrators protested the peripheral involvement of school people in the programme's development; and alleged poor technical quality of the tests, as well as unethical behaviour of the Department in releasing the scores. Some teachers complained about non-involvement, that the tests were

not valid, that it was not proper to hold them responsible for student progress, and that the lack of individual score information for each student made the data useless to them. The following quote captures some of the feeling that the score-reports engendered:

> The (program) is really politics masquerading as research. Promise after promise has been broken....Plans have been dictated and changed by the legislature....It is not an operational purpose of the assessment project to improve instruction by identifying promising practices....The conclusions were written before the project was undertaken....Educators at the district level have not been included in designing the tests and are not included in plans for the development of future tests (Grosse Pointe Public Schools Administration, 1970 as quoted in Kearney and Huyser, p. 47).

During the second year of the programme, the Department responded to some of the pressures generated by the release of the first-year results. Initially, the Section 1 background information was to be collected anonymously. Later, the collection of all such data was stopped. There was a restatement of programme purpose. The programme was announced as not only aimed at aiding decision-makers at the state level, but also as aimed at providing useful data to decision-makers in the schools. Individual pupil rather than group scores were to be reported thereafter.

Despite the Department's attempts to respond to criticisms, the assessment was negatively received in many quarters the second year also. Some teachers maintained that the test results were still of no help in diagnosing children's learning problems, particularly because the tests were not based on Michigan curricula.

What had originally begun as an effort to rationalize decision-making at the state level in reallocating resources by linking them with school outcomes had developed into an effort to diagnose individual learning problems. The assessment's relevance to state policy seemed less immediate. The assessment's decrease in relevance to state policy appeared to undermine the original rationale for the programme (Murphy and Cohen, 1974).

A new design for the programme was announced early in 1972. The new design stated that after the 1973 assessment, the original norm-references tests were to be replaced by objective-referenced tests. A task force composed of Michigan educators, students, and lay citizens had been appointed by the State Board of Education in 1970 as part of the original three-year planning and development period and had been charged with the task of developing statewide goals. After 25 public meetings and more than a year, the Common Goals of Michigan Education were adopted. These goals were translated into specific objectives by several groups composed mostly of professional educators. The objectives were directed at "minimum learning expectations" for all children. The "minimum learning

expectations" had already been generated by nine "Reference Groups". (The political issue of using "minimum learning expectations" was explored further by Murphy and Cohen, 1974).

Commissions composed of teachers, administrators, curriculum specialists, School Board members, parents, and other citizens were formed to review the minimum developed by the "Reference Groups". After reaching concensus on a number of objectives, test development in reading and mathematics began. A year was spent in item writing and pre-testing. Tests in reading and mathematics were administered to all 4th and 7th graders in the fall of 1973. The results were sent to each district.

Many teachers felt that the information provided was better than that available from the old norm-referenced tests, but they still objected to the implication that teachers were primarily accountable for student performance. On behalf of a reticent membership, in January, 1974, the Michigan Education Association and the National Education Association (both professional organizations for teachers) brought in a "blue ribbon" panel of experts to evaluate the Michigan Educational Assessment Program. The panel stated that the Michigan model had good features as well as bad. In criticism they stated :

> The most serious flaws are in the assessment component; it is too narrow in scope to serve as a state needs assessment and it has been implemented on an every-pupil basis without technical or utility justifications. Overall, there is no clear evidence that state and local decision-making is being served by the accountability model (House, Rivers, and Stufflebeam, 1974).

Their report went on to state :

> The accountability and assessment systems also risk substituting state objectives and curricula for local ones, not a reasonable educational trade-off in our opinion (House, Rivers, and Stuffebeam, 1974).

The "blue ribbon" panel suggested more gradual development, matrix sampling, putting the assessment on a voluntary rather than mandatory basis, and placing more emphasis on the development of school-based evaluation as possible ways to improve the assessment's quality and utility. Three staff members in the Michigan Department of Education responded in print to the "blue ribbon" panel's evaluation (Kearney et al., 1974).

In a speech entitled "Limitations and Problems in Measurement in Mandated Assessment Programs", presented at the Large School Systems Invitational Conference on Measurement in Education, Robert Ebel in referring to the "blue ribbon" panel's report stated, "It is a readable but superficial and biased assessment of a pioneering and generally sound program of state assessment" (Ebel, 1974).

In their analysis of the assessment in Michigan, Murphy and Cohen (1974) assessed the outcomes of the effort in the following way:

One hope among state officials was that the collection and publication of data on school resources and outcomes would provide a scientific basis for allocating more resources to poor districts. That notion has more or less fallen of its own weight, for the assessment - like most other large-scale research on school effects - turned up little relationship between school resources and student performance. If anything, the assessment provided comfort and ammunition for the enemies of liberal reform.

Another hope was that the assessment program could provide an evaluation of the schools that distinguished effective from ineffective programs. This way taxpayers could see whether they were getting their money's worth. That hope also has been dashed. The data were not really appropriate - though they were used anyway - and the state did its best to avoid this problem by disavowing any evaluative purposes.

Another expectation was that better information would make decisions better by making them more rational. Michigan educators don't seem to have any more idea of what a "better" decision is than we do, but it would be hard to find anyone in Michigan who thought that the assessment had made things more rational. The experiment has generated a steady background of political conflict and confusion, intermittently punctuated by major explosions. One consequence has been to force several major revisions in the assessment program. The most important of these, the switch to objective-referenced tests, has posed an entirely new set of problems and generated a mountain of novel informations whose meaning and relevance to policy is unclear (Murphy and Cohen, pp. 72-73).

Although it may be too early for any definitive judgments about assessment in Michigan, it seems that any large-scale assessment programme might expect to encounter many problems. It should be pointed out that this brief case study has been presented, not with the purpose of evaluating the Michigan Educational Assessment Program but to describe the negotiations and commissioning of an important large-scale evaluation activity. The problems arising in this example might be instructive to individuals who would desire or are obligated to initiate such an evaluation activity.

REFERENCES/SOURCES

This case study is heavily based on those sources with an asterisk(*) in front.

* BERKOWITZ, Norman, Deputy Secretary of State, State of Michigan. Telephone conversations, October and November, 1974.

EBEL, Robert. Limitations and problems in measurement in mandated assessment programs. Paper presented at the Large School Systems Invitational Conference on Measurement in Education in Cincinnati, May 1974.

HOUSE, Ernest, Rivers, Wendell, and Stufflebeam, Daniel. Assessment of Michigan's accountability system. Teacher's Voice Supplement, Michigan Education Association, April 22, 1974.

HOUSE, Ernest R., Rivers, Wendell, and Stufflebeam, Daniel L. A counter-response to Kearney, Donovan, and Fisher. Phi Delta Kappan, September 1974, LVI (1).

* KEARNEY, C. Philip. The politics of educational assessment in Michigan. Paper presented at the annual AERA convention held in Minneapolis, 1970.

* KEARNEY, C. Philip, Associate Superintendent for Research, Michigan Department of Education. Telephone conversations, October and November, 1974.

KEARNEY, C. Philip, et al. Improved information for educational decision-making: The Michigan assessment program. Administrator's Notebook, February 1970, XVIII(6), p. 2.

KEARNEY, C. Philip, and Huyser, Robert J. The politics of reporting results. In E. House (Ed.), School Evaluation: The Politics and Process. Berkeley, California: McCutchan, 1973, p. 48.

KEARNEY, C. Philip, Donovan, David L., and Fisher, Thomas H. In defense of Michigan's accountability program. Phi Delta Kappan, September 1974, LVI(1).

MICHIGAN Department of Education. Purposes and procedures of the Michigan assessment of education. Assessment Report No. 1. Lansing, Michigan, 1969. In C. Philip Kearney and Robert J. Huyser, The politics of reporting results. In E. House (Ed.), School Evaluation: The Politics and Process. Berkeley, California: McCutchan, 1973, p. 48.

* MURPHY, Jerome T. and Cohen, David K. Accountability in education - the Michigan experience. The Public Interest, Summer 1974, 36, p. 56.

* POLLEY, Ira, Director of Admissions and Scholarships, Michigan State University. Telephone conversations, October and November, 1974.

THOMAS, J. Alan. School finance and education Opportunity in Michigan. Lansing, Michigan: Michigan Department of Education, 1968.

5. A CASE STUDY OF INFORMATION RETRIEVAL TELEVISION

Information Retrieval Television, dubbed IRTV, was the birth child of the Northern Electric Company of Canada, presently amalgamated with Bell Canada Limited. The concept was originally worked out by these public utility companies and presented to the Public School Board of Ottawa and the Ottawa Collegiate Institute Board. As a summary Statement of the system, John Twomey wrote:

> The system...was to involve a central library of video tapes and films which may be called up via telephone from any classroom in the participating schools linked to the library-control centre by coaxial cable. Transmission of the film or taped material was to be made from the centre via cable to a TV receiver in the classroom (Twomey, 1971, p. 2).

This automated-booking system was to involve a collection of elements: classroom TV sets, film and TV programmes, librarians, telephones, and machines, all interdependent. The main aim of the project was to provide teachers with the opportunity to use programmes in the class when and where the programmes would be of most benefit to the students. The only access teachers previously had to films was through regular educational-television programming and through conventional film distribution methods. The report of the IRTV experiment, itself, was the two-year evaluation study conducted by the Ontario Institute for Studies in Education (OISE).

Initial Contact

As the concept of an on-line booking system emerged, Bell Canada, Northern Electric, and the two Ottawa Boards wanted to put this idea to a proper test. All four parties, however, had a vested interest in the system, therefore they decided that an outside party was needed. OISE was a new provincial institution of two years with generous support received in block grants from the Ontario government. In the early part of 1967, the director of OISE received a letter from M.C. Billowes of Northern Electric concerning an in-house proposal for an on-line retrieval system. The director passed the proposal on to David S. Abbey, then Assistant Coordinator of Research at OISE. Abbey realized the potential benefits of OISE's possible involvement in the IRTV project. OISE was a new institution designed specifically to conduct studies in education. The prospect of a research investigation conducted in conjunction with industry looked to be a promising endeavour and could provide a basis for good public relations with industry and business. In addition, the project was personally

attractive to Abbey who was interested in media research in general. The programme would entail many man hours of work. It could also be used as a training project for graduate students. At this time no formal request for an evaluation study had been submitted.

Bell, Northern Electric, and the Ottawa Boards wanted to get the project under way. This involved a commitment from OISE which was not forthcoming. Abbey so far had been the only person at OISE interested in the project and he was willing to work at it only on a personal basis. A commitment from OISE could not be obtained unless some person or department was willing to take on the project. A meeting was held at OISE in the spring of 1967. The proposal was presented, and research prospects on the IRTV system were discussed. No follow up of the meeting occurred and no proposals were forthcoming for a research design. Therefore, Abbey approached Leslie McLean, the chairman of the Department of Computer Applications at OISE and suggested that this project would be one his department could consider. It was apparent that if the Institute was to present a good face to the public then an affirmative answer had to be given; otherwise the proposal would have to be turned down as a project for OISE. Therefore, McLean, rallying forces and with high hopes, accepted the proposal. In the negotiations which followed all parties tried to delineate the terms of a working agreement. One problem was how to go about formally inviting OISE to join the IRTV project.

It was decided that OISE should be formally invited to participate by the Ottawa Boards. The formal request came on August 2, 1967, from J.W. Grace, Chairman, Collegiate Board of Education and K. A. Murchison, Chairman, Public School Board of Education of the city of Ottawa. It was received by R. W. B. Jackson, Director of the Ontario Institute for Studies in Education.

During the autumn of 1967 members of the Department of Computer Applications and Curriculum at OISE held meetings to determine the nature of OISE's involvement in the programme and to develop a research design for the IRTV project. The major concerns of the Institute were presented by Abbey: 1) that the audio-visual material should be evaluated before it was introduced into the classroom through IRTV, 2) the present use of audio-visual instructional-devices within the schools participating in the project should be assessed, 3) careful consideration should be given to the cataloguing and indexing of the software so its relevance to the present curricula is determined, and 4) the measurement of the ways in which the material is incorporated into the classroom and the effects it would have on the teacher and students should be given careful attention (Twomey, 1971, p.4). These concerns were presented to John H. M. Andrews, OISE Coordinator of Research by Abbey in a letter dated

December 11, 1967. The letter also indicated OISE's strong interest in developing the IRTV project and committed the Institute to the project.

Early Negotiations

Negotiations over funding were not a major issue of the discussions. OISE was receiving block grants which were unallocated. Thus it was easy for a department chairman to take on a project such as IRTV and get a budget authorized. McLean simply added the project to the already anticipated work for the new budget year of his department. (In such a situation, where the parties requesting evaluation assistance do not provide funding, an unusual non-hierarchial organization may develop). No money was exchanged between any of the parties. Bell Canada and Northern Electric provided the hardware. The Ottawa Boards had a sizable expenditure in the library and classroom monitoring system, while OISE contributed the evaluation endeavour and - because no one else took the initiative to do it - the production of a catalogue.

With sources of funding decided, the focus turned to the consideration of issues to be addressed in the study. Bell and Northern Electric were not specific as to expectations they might have from the evaluation study. They were mainly concerned that a thorough evaluation be done, anticipating that only then could a sound judgment be made as to whether the system was of enough benefit to warrant expansion and dissemination.

The Ottawa Boards were interested in whether or not the system would work. The questions posed by the Ottawa education authorities focused on 1) a cost evaluation of the project 2) the educational benefits accrued by children exposed to visual aides at suitable times in their instructional programme 3) and the effects this system would have on the role of teachers. There were different reactions among board members, of course, with some determined to have it work at all cost, and others not at all interested in the system. OISE had yet to define its position. McLean did not want to make the project a study of educational television, but rather a study of IRTV. This was a new programme and not everyone, including the developers, knew what was going to happen. The evaluation could easily have turned into a test of the hardware. However, OISE was concerned with testing the medium as well. No one as yet knew what the parameters of the system were going to be.

But before negotiations about the actual evaluation were under way, problems involving working arrangements and the delegation of authority had to be settled. The early months of 1968 seemed to be taken up with public relations aspects of the project. All parties

were contributing much in terms of time, personnel, and money. Problems arose because of the different interests each party had in the project. There was considerable ego-involvement on all sides; Bell Canada and Northern Electric were big companies, prominent in the public eye. OISE was a new institution struggling to establish a reputation and with a mandate to do research. The Ottawa Boards were introducing an instructional system which would hopefully benefit their curricula and gain them a reputation as innovators in the field.

Communication between the parties was hampered by the locations of their head offices. Bell policy-makers were situated in Montreal, OISE had its centre in Toronto, whereas the project operation was in Ottawa. Therefore, when letters and telephone conversations were insufficient, meetings had to be juggled among the three locations.

Another problem that emerged as the study got under way concerned the regulations under which each of the parties was operating. The Ottawa School Boards were under provincial jurisdiction while Bell and Northern Electric were private companies - yet both possessed federal charters. Therefore the different legal regulations by which they were bound placed additional restrictions on the system. For example, a technician from Bell could not take a reel of tape and place it on the school's equipment. He had to give it to a school technician even though the Bell man was to operate the equipment. The tape was school property. Such roundabout operations proved to be added burden in terms of cost, personnel, technical difficulties, and personal problems among the parties.

One of the major problems visible from the onset of the project was the hierarchical organization within the agencies involved. This was particularly a problem in the case of the Ottawa Boards. In his December 11, 1967 letter, Dr. Andrews had requested that the two Ottawa Boards select a person with the authority to act for both and insure clear communication. As of September 9, 1968 no such person had been appointed and the person "standing-in" was doing so only part-time. There was little doubt that the project required a full time representative in order to insure close cooperation and communication. This slowness on the part of the Ottawa Board greatly hindered negotiations in the project.

For example, arrangements had to be made about the allocation of the materials once the project was completed. It was decided that the coaxial cables installed were to be left operational and that they would become the property of the Ottawa Public School Board upon completion of the study.

Questions about reporting evaluation findings also had to be negotiated. OISE maintained that it would not submit the content of the evaluation report to the coordinating committee for editing or decisions as to dissemination. In the tradition of academic freedom,

they demanded the right to report the results as they saw them. After "many bottles of scotch and an arm wrestling match in an Italian restaurant" it was agreed that OISE would compile the final evaluation report but would allow the other parties to view it before its release to the public.

The question of how information on the development of the project was to be given via television and newspaper articles to the public was troublesome. It was decided that no information would be disseminated without the consent of all the parties. The main reason was that with five parties involved in the project there was a very real problem of issuing separate and possibly conflicting statements on the IRTV project. Also, since the attitude of the public toward the IRTV project was one issue under investigation there was an experimental need to control the information publicized. Mrs.Eleanor Senter of Bell Telephone's publication department was placed in charge of releasing information to the news media. The first release came on May 31, 1968 having been circulated to all parties and approved.

The initial establishment of communication was a problem. A coordinating committee was set up at the request of McLean. The purpose of the committee was to settle disputes among the major parties involved. The committee probably was the "saving grace" of the ensuing negotiations. Any actions or decisions to be made on the IRTV project had to be passed by this committee.

Most of the evaluation "contract" issues confronting the parties were resolved, not by setting up rules but by indicating who would have jurisdiction to make decisions.

Selection of Schools

By January 1968, the agreement concerning selection of schools had been completed between the Ottawa Boards and Bell Canada. These schools were to be connected by coaxial cable. OISE was not at all involved in the selection of the schools. The schools selected for the study included one high school, an intermediate school, and two elementary schools (in 1970 one more public school was connected). A corresponding set of schools was selected by the Ottawa Boards as a contrast group. However, the high school selected for the study was a new one. The other high school in the contrast group was an established school known for its academic standing. Some staff members of this school felt that all new and interesting projects were being given to the other school. They felt slighted, were angry, and perhaps were determined to help the project fail.

The IRTV schools served students from a lower economic community than that served by the contrast group. This was due to the fact that an initial decision had been made to select the high school and take the "feeder" schools nearest it into the project.

OISE had the prespecified role of evaluation. However, they did not want the evaluation to become a study of the teaching potential of TV. From the numerous discussions occurring during the first quarter of 1968, the design of a formal research study for the evaluation had to be developed. At a meeting held in Ottawa on March 6, 1968, Abbey set out several important research conditions for the evaluation of the project:

1. An automated recording device will be provided to indicate channel usage in the system.
2. All participating schools will have access to the full complement of A/V materials normally available to students of the two Boards.
3. Any additional material acquired for purposes of this study will be included in this "common library".
4. A selection of films and TV tapes in the IRTV library and available for experimental schools will be duplicated for use in the common library.
5. All schools (including IRTV schools) will have access to the "common library" through normal ordering procedures.
6. The creation of the catalogue for the A/V and IRTV libraries will be the responsibility of the Ottawa School Boards in consultation with OISE (Twomey, 1971, p. 12).

On April 22, 1968, R. W. B. Jackson of OISE sent a letter to Bell Telephone, Northern Electric, and the Ottawa School Boards confirming OISE's commitment to the project and specifying the details of the project and a time schedule for activities. McLean was formally appointed as the person responsible for the evaluation. On May 23, 1968, he sent a letter to the principals of the schools explaining OISE's role in the project. This letter was thought to be vital in facilitating later negotiations.

In an earlier correspondence McLean had written:

> The general procedure will be to test several hypotheses related to the assumptions that easy and rapid access to a large amount of audio-visual material could make a significant difference in the quality of classroom instruction and learning" (Twomey, 1971, p. 10).

Even though the purpose of the evaluation study was thus well documented recurring debates ensued. Not all people involved in the study saw the letter; some who did forgot it. The cry that no tests were being given to the students was often encountered. Some of the areas to be examined in the study included traffic data on the system, cognitive educational variables, noncognitive variables, materials requested by phone, social pressures, reasons for using IRTV materials, instructional innovation, teacher role, administrative structure, cataloguing and indexing, educational TV, and use of the IRTV system.

Teacher Contact

The early planning of the programme anticipated a phase-in-period for the fall of 1968 with the two year evaluation extending from January 1, 1969. However, the actual phase-in-period was postponed until December 1, 1968. Expectation among the staff of the various experimental schools was rising. The first meeting with the IRTV school personnel was held on October 1. Representatives from the two Ottawa Boards, Northern Electric, Bell Canada, and OISE were present. Teachers were encouraged to ask questions and were informed of the developments. The enthusiasm for the programme was not as great as had been anticipated. This could have been a reaction to the pressures of a new school year which had just started in September.

On October 2, the same people with the exception of the representatives from Bell Canada and Northern Electric spoke to the staff of the contrast high school. The reception was somewhat cold. On October 3, another meeting was held with the staff of the three IRTV public schools and the three public schools of the contrast group.

Research and Data Collection

Research proposals were accepted from individuals interested in the IRTV project. These were approved by the OISE Research Review Board. Budgets had been allocated in July, 1968 for the first year of the project. Some of the research undertaken at various points in the IRTV project included:

"The Parent and Student Attitude Study" by David S. Abbey

"Information Retrieval Television: Teacher Attitudes" by M. W. Wahlstrom.

"Student Questionnaire Report" by G. Harry McLaughlin.

"Report on Discussions with Children in Kindergarten to Grade 4 about Television Programmes and IRTV" by the Carota Family.

"Learning from Color versus Black and White: A Comparison Using IRTV and Film" by M. W. Wahlstrom.

"Instructional Climate in Two Ottawa High Schools" by N. W. Wahlstrom.

"Head-end Operating Procedures" by Dennis Dibb.

"A Computerized On-Line Booking System" by M. W. Wahlstrom, G. H. McLaughlin, R. Dinning and Valerie Grant.

"Report on Usage of Material in the Ottawa Board of Education Audio-Visual Centre" by Merlyn Gaynor and Charlotte Reid.

The plan for the actual evaluation study was informal (some said "loose"). In fact it was hardly designed at all but rather was worked out as the system came into operation. This approach was defended by McLean because the system itself was evolving and changing. Instructional objectives had been stated. The purposes even of the evaluation were still emerging.

In a decision he later regretted McLean selected a businessman to be OISE coordinator of the IRTV project. According to McLean, this man did not perform effectively. In addition the idea of using a person only with a business perspective to coordinate OISE's relations with Bell and Northern Electric was unworkable because he could not understand the educational concerns of OISE. In 1970 G. Harry McLaughlin joined the IRTV project as Coordinator "to salvage the operations". McLaughling had worked with the National Aeronautics and Space Administration and with the Department of Public Health and Welfare in the United States. He was familiar with the major concerns of big business, yet he had academic experience too. His chief advice to other education investigators seeking a joint project with industry is to "select an academic who knows how to yell at people".

Data for the final report of the project were collected with questionnaires by a trained interview or questionnaires filled out by the respondents themselves. There were no severe problems concerning access to information although Bell Canada and Northern Electric did consider cost information as propriety. The questionnaires were distributed in the following sequence: in November 1968, 100 members of the general public and 500 parents of children attending IRTV and comparison schools were interviewed concerning their awareness of IRTV. Six months later in 1968, 318 of the originally sampled parents were reinterviewed along with 204 of their children. Seventy-five members of the general public were reinterviewed along with 25 people who had never before been questioned. In November 1968, 46 IRTV teachers and 20 comparison teachers filled out a questionnaire concerning their attitudes towards new educational media. In May 1969, all IRTV teachers filled out an abbreviated version of this questionnaire. The final questionnaire concerning attitudes towards the use of television and other audio-visual media was completed by 81 IRTV teachers, 82 comparison teachers and 31 administrators in February 1970.

In June 1970, all 1831 available students in Grades 9 and 11 English students and Grades 10 and 12 science students as well as 38 of their teachers in IRTV and comparison schools were provided with a questionnaire on class activities. Also in 1970, two groups of Grade 7 students totaling 239 participated in an experiment to determine the value of color television. In January 1971 all students from Grade 5 to Grade 12 were given six different questionnaire forms.

Ms. Charlotte Reed, an OISE Research Assistant stationed in Ottawa, interviewed 31 teachers in December 1969. In the spring of 1970, 41 IRTV teachers who had not completed the 1970 questionnaires

were interviewed. Some 419 children in Kindergarten through Grade 4 were interviewed to determine their opinions of IRTV. These interviews were conducted by six children aged 10 to 17 years under the supervision of their parents who were trained interviewers and teachers.

Additional data were collected on IRTV bookings and transmissions from its inception until March 1971. All orders of audio-visual materials made by IRTV and comparison schools were recorded from September 1967 to December 1970. Records were kept of the IRTV and audio-visual library holdings.

Problems along the Way

One of the major issues that arose regarding the project was the inability of the Ottawa School Boards to meet the research conditions set out by Abbey. The Ottawa School Boards were not able to complete the film library. By the spring of 1968 it was evident that the previewing and processing of the programme software was a massive endeavour. OISE volunteered assistance. Ms. Reed moved into the IRTV high school and organized the previewing of films.

Because of the IRTV project and through the efforts of the "businessman" hired by McLean, many items were given to the library which previously had not been available to the schools. While the final library holdings did contain much that was not useful, on the whole it contained proportionally more good material than had previously been available to the schools. When establishing such a library system, the question of quality of material was side-stepped in order to obtain enough volume so that a significant number of the requests made by the teachers could be fulfilled.

The summer months of 1968 were to be used to prepare the library for its opening in the fall. However, the slow rate of growth of the library became a major concern. The phase-in-period was postponed until December 1968 and at that time only 642 programmes were previewed, many less than the 2,000 titles anticipated by Abbey. Since Abbey had stipulated that comparison schools were to have access to the films used in the IRTV library all IRTV materials had to be duplicated. The team assigned to this aspect of the project was greatly overworked and lacking some of the expertise needed for ordering, receiving, previewing, and abstracting film materials.

As the IRTV system approached its phase-in-period it became evident that no consideration had been given to the preparation of a catalogue. OISE took on this activity although initially it was the responsibility of the Ottawa School Boards. The interim catalogue was distributed in November 1968. In March 1969 the first cross-referenced catalogue was prepared with additional editions made in June 1970 but not distributed to the schools until October 1970 due to editing corrections.

Another concern was the use made of IRTV by the teacher. There was, of course, no experimental control over how the teachers used the system. It was found to be used as a supplement to the main teaching activity, seldom used as a major instructional device in the school programme. Results indicated that IRTV was used by 40 per cent of the teachers to teach the main lesson. Thirty per cent of the teachers used it to summarize or review material; only 12 per cent said they used it to introduce a topic. It was felt by OISE that the Ottawa School Boards could have influenced teachers' use of the system but such was not done. This was believed to be a major cause of teacher apathy.

There was also a lack of communication between the teachers and OISE. Although a research assistant had been stationed in Ottawa more direct contact with the institution would have been desirable. Many research opportunities were missed because of this lack of communication. In addition many teachers did not complete the questionnaires. This was apparent especially in the comparison high school whose staff had already expressed resentment towards the project.

One of the misunderstandings between the Ottawa School Boards and OISE that became apparent was about student access to the system. In his initial contact with the school boards, Abbey had assumed that students would have access. This condition had not been the expectation of the school boards, and students were seldom allowed to use the system.

The number of manhours and personnel needed for the project was greatly underestimated. Also, a large amount of software had to be purchased by the school system. No one had worked out the costs of duplicating the library. This proved to be an enormous cost both in terms of money and the time and personnel required to complete the task.

With regard to the interaction among the parties involved, the contingencies were not worked out as to what would happen where one person or party might fail to follow through on their tasks. The problems which resulted are exemplified in the Ottawa School Boards' slowness in appointing a person to represent them jointly. The inefficiency in the previewing and processing of the films and the catalogue were the results of a party's inability to fulfill contractual obligations, thus adding an additional burden onto the other parties involved. Also the lack of contact between OISE and the teachers was in part due to the non-interfering policy of the Ottawa School Boards which may have affected the actual use of the IRTV system. It is not obvious that additional pre-arrangements or more authoritarian control would have made the evaluation more effective, but it would probably have made the working conditions more suitable for the evaluators.

Results

The main aim of the evaluation team was to view the IRTV system in terms of its efficiency as a means for delivering audio-visual material. Attitudes toward the system were to be gathered from teachers, administrators, students, parents, and the general public. A cost analysis comparing IRTV with other methods for delivering audio-visual materials was to be prepared.

The results of the IRTV project indicated that IRTV was used mainly as a supplement to other instructional techniques. In addition the cost of the IRTV greatly outweighed its actual use in the classroom. Programmes available for the system did not cover all areas nor were they particularly good as teaching devices. The IRTV teachers reported that they tried to engage their students actively before, during, and after viewing films and television programmes, more than did the comparison teachers. The average IRTV student was likely to have a greater variety of types of assignments than the comparison student.

The major reason for the IRTV system not providing enough benefits to justify its costs was due to the lack of good teaching programmes presently available in the system and the nature of the organizational aspects of the present school system. The evaluation report maintained that, given the teacher-dominated role in classroom activities, the school environment did not as yet appear ready for the IRTV system. There was greater use of the film and video tape materials by IRTV teachers. This had been anticipated. The average student in a comparison school viewed 1.03 programmes per week while a student in an IRTV school viewed 2.11 programmes. One complaint voiced by the teachers was the lack of sufficient programmes in a particular field.

The anticipated general advantages of IRTV over film were confirmed that: 1) it insures that a teacher will be able to obtain a programme within a very short period of time after requesting it; 2) it facilitates viewing of programmes by individuals or a small group of students; 3) it does not bind a class to an advanced plan since IRTV programmes could be obtained on demand.

Reporting

The report was written primarily for decision-makers, e.g., trustees and administrators. Secondly it was directed to teachers and finally for everybody else. It had not been designed particularly for use by researchers.

The issue of the control over the final report had been discussed during January of 1968. It was decided, as has been mentioned, that OISE would have complete control over what and how results would be

reported. However, dissatisfaction arose when the final report was submitted to the sponsors, especially Bell Telephone. Bell representatives felt that the OISE report was slanted toward the evaluation of film which was not the issue under study. They pointed out that IRTV represented more than just films. They also questioned the statement in the report that "only 33 per cent of the teachers used the IRTV system". Bell felt that this was a significant difference from none at all. However, Bell had no control over the report at this point and OISE printed the report as it stood, without offering sponsors an opportunity to include a rebuttal or disclaimer.

Conditions Insufficiently Considered and Advice for Future Research

When working arrangements are set out between the parties involved in an evaluation and when the actual programme is undertaken it is, of course, impossible to foresee all the possible problems that may occur. The actual use of the IRTV system probably had been insufficiently considered; consequently the evaluation design was probably too indefinite.

If the study was to be done again, it was felt by the evaluators that a more controlled field experiment would be desirable. Also rapport with the people using the materials, e.g., teachers, students, principals, seemed to be needed at the beginning of the project. OISE would perhaps have moved closer to the setting and possibly would have generated an experimental research project, providing a trusting relationship could have been established.

REFERENCES/SOURCES

ABBEY, David S., Assistant Coordinator of Research, Ontario Institute for Studies in Education. Telephone interview, October 1974.

BURWELL, J. Hugh, Assistant Superintendent of Programme and Planning, Ottawa Board of Education. Written communication, September 1974.

FROHLOFF, Ernest C., Headquarters Marketing-Special Services, Bell Canada Limited. Telephone Interview, October 1974.

GAILITRS, Maris M. The costs of information retrieval television: A case study in the cost-effectiveness of educational media. Occasional Paper No. 12. The Ontario Institute for Studies in Education, Toronto, 1972.

MCLAUGHLIN, G. Harry. Educational television on demand: An evaluation of the Ottawa IRTV experiment. Occasional Paper No. 11. The Ontario Institute for Studies in Education, Toronto, 1972.

MCLAUGHLIN, G. Harry, Coordinator of IRTV Project, Ontario Institute for Studies in Education. Telephone Interview, November 1974.

MCLEAN, Leslie D., IRTV Project Director; Chairman, Department of Computer Applications, Ontario Institute for Studies in Education. Personal interview and telephone conversation, August-September 1974.

MINUTES of IRTV, Coordinating Committee Meetings. The Ontario Institute for Studies in Education, September 1968, April 1971.

TWOMEY, John. Early history of the information retrieval project. The Ontario Institute for Studies in Education, Department of Computer Applications, September 1971.

WAHLSTROM, Merlin W. Co-director of IRTV Project, Ontario Institute for Studies in Education. Telephone interview, November 1974.

6. EVALUATION AS A BASE FOR SCHOOL REFORM
- THE CASE OF SIA IN SWEDEN -

The Kilborn-Lundgren evaluation study of 1974 was a brief and intensive review of experiments which had been conducted in ten Swedish schools. The experiments had been initiated by SIA, a special parliamentary committee on the internal operations of the schools. The experiments were efforts to try out locally developed solutions to the problems of teaching the national curriculum to children with various learning handicaps.

The SIA committee had been assigned the responsibility of designing a broad new reform programme for the Swedish schools. Evaluation of the schools had shown that some differences in achievement were still attributable to social factors in spite of the vigorous reform efforts of the early sixties to eliminate differences in educational opportunities. The experiments were based on the same ideology as the original reforms. Directives to and from the committee emphasized the need to find means of reallocating existing resources in ways which would give greater assistance to handicapped students, including those with social and intellectual handicaps.

It was anticipated that the SIA committee would submit a legislative proposal outlining needed changes in existing laws and regulations. The committee's task was to consider long-range as well as the immediate implications of any inadequacies or deficiencies in existing programmes and to develop an ameliorative programme of pedagogical changes and compensatory measures that would help make the school experience more stimulating and meaningful for students having difficulties with school work. In addition the committee was to identify problem areas and questions requiring further research and investigation.

The experiments in the ten schools were initiated in an attempt to investigate alternative ways of using personnel resources. Each interested district proposed a set of innovative procedures (within certain guidelines established by SIA) for organizing and teaching classes containing some of these handicapped children. Ten schools from various parts of the country were selected for participation.

In Sweden, the assignment of teachers, time tables, and the content of the curriculum is prescribed by the National Board of Education. The experiments permitted some loosening of the tight requirements on staffing and organization of classes. These experiments were intended to be carried out without altering the existing

curriculum plan and without any form of ability grouping. The experimental treatment was the loosening of organizational frames as well as some decentralization of decision-making.

During the two years of the experiment which began in 1971, the ten schools submitted periodic reports to the committee. Some of the reports were detailed and descriptive, others were quite vague. Almost without exception they were focused on organizational problems. None of them explained the pedagogical activities, that is, what the teachers did in the new arrangements. Hence, what some considered to be the most essential information from the experiments was not available to the committee.

Early in the autumn of 1973 - with the two years at a close, with the final reports from the schools coming in (with less than the information needed), and with the dates for submitting new legislative bills drawing near - the committee looked for help. The secretary asked Ulf Lundgren, an experienced educational researcher from the University of Gothenburg, and his colleague, Wiggo Kilborn, a mathematics curriculum developer, to conduct an evaluation.

Within the committee there were differences of opinion as to which questions the evaluators should focus on. The secretary favoured placing greatest attention on the organizational rearrangements. But some of the committee members wanted the evaluators to focus on pedagogical and sociological issues, such as whether or not slow learners could effectively rejoin the main group after being isolated for days or for months in a remedial group.

Lundgren and Kilborn proposed to review all reports; to visit all schools (some of the experimental arrangements were continuing that autumn even though the experiment had officially ended); to interview teachers and others about the teaching, learning and organization; and to report back by the first of the year. Informally, often in conversations between one or two committee members and the evaluators, the probe was discussed. Within a few weeks after the initial invitation, the committee agreed to support the inquiry largely as Lundgren and Kilborn had proposed it. During the months of October, November, and December, that rather informally negotiated plan was carried out.

The actual writing of the final evaluation report was coordinated with the writing of SIA's legislative proposal. The evaluators' aim was to provide SIA with more information about the effectiveness of the various innovations and the generalizability of the knowledge gained from the experiments. One of the main recommendations emerging was to permit all schools to follow the more flexible directives which had been given to the experimental schools.

The principal mode of inquiry utilized by the evaluators involved interviewing teachers, pupils, administrators and local board members. In conducting these interviews, Kilborn and Lundgren attempted to discover the uniqueness of the approach used at each school and what there was about that particular approach which might be generalized to other Swedish schools, especially in various urban and rural contexts. The basic question they sought to answer at each school was, "What possible consequences would this model have if it were used under the different circumstances of another school"?

Kilborn and Lundgren viewed evaluation as a process of comparing expected results with observed results. They reported that one of the major difficulties they experienced in attempting to evaluate the various experimental approaches was that in most cases the expected results were not defined. They had to be inferred. In the documents from most schools and even in the documents of the SIA committee, they found the pedagogical ideas not sufficiently explicit. It was difficult to answer a straight question as to why a certain organizational arrangement was expected to benefit a learner.

The evaluators discovered that in many cases the guidelines which SIA had established for the conduct of the experiments had not been followed. Various schools had interpreted the guidelines in different ways. In their attempts to manage discipline problems and to be innovative the schools had taken unauthorized initiative. The directives regarding the amount of time which could be allocated to various subjects were often compromised. For example, hours designated for home economics had been converted into hours in physics and mathematics. (Teachers of practical subjects were especially interested in this part of the evaluation. An issue of one of their journals was devoted entirely to the study).

The secretary (executive officer) and each member of the SIA committee were eager to see if this decentralization could be accomplished within the prescribed curriculum particularly in terms of the schedule and regulations for instruction. The interest of the public also was considerably aroused. Not only the professional newspapers of the teacher unions but the national press too followed the progress of the experiments. The experiments and reform efforts were seen as relevant to the growing risk of teacher unemployment. It was realized that some schools had been using allowances for remedial teaching, in effect, to reduce class size and thus to maintain a larger regular teaching staff. If the experiments turned out favorably this would support arguments for allocating more public resources to the schools for lowering class size, without drawing upon the resources needed for remedial teaching.

The data from the schools' reports were insufficient for dealing with this important public issue Furthermore, some of the schools apparently had been able to obtain some reallocation of resources from local authorities - but this information was not included in the reports to SIA. Some of the parliamentary members, those with strong convictions about the anti-democratic effects of sub-grouping in the classroom, wanted to know more about the resources being used for lowering class size. Several research studies related to the same basic problems had been commenced by SIA, but they were not expected to shed light on the issues of immediate concern. More information was needed about the results of these ten experiments.

Another example of bypassed guidelines was the grouping of students. In some schools all extra resources were used to lower class size. This meant that some teachers trained to teach handicapped children were assigned classes of normal children. This was directly contrary to the basic intent of the SIA committee, to find better ways of allocating resources to children with learning difficulties. In some of the experimental schools there was a high degree of ability grouping, a practice allowed neither by the national curriculum plan nor by the guidelines for the experiment.

In their conclusions the evaluators reported that the loosening up of the rigid staffing and grouping patterns seemed to have worked successfully, certainly to the satisfaction of the people involved. They cautioned that more effective guidelines or regulations would be needed, and probably continuous evaluation. They concluded that if students with special problems were to be allotted more instructional resources new types of steering instruments would have to be developed and the curriculum guide would have to be revised. Overall, the evaluators considered the experimental results as positive, but they suggested that further experimentation was necessary. They warned that too much decentralization can mean that the schools will develop in directions contrary to those now set for Swedish education. They emphasized the need to develop new courses in the schools and new instruments for governing remedial teaching assignments. This part of their evaluation report proved to be very controversial on the political front.

The evaluators recommended that the rules and regulations governing staffing, grouping, and scheduling be broadened or changed to permit greater flexibility. However, they were quick to point out that this change would have to be accompanied by a more overriding and continuous evaluation. They expressed confidence that this could and should be, at least in part, a decentralized form of evaluation in which individual schools would have the opportunity of evaluating the effects of different approaches by themselves.

In addition, Lundgren and Kilborn declared that experimental efforts needed to be continued. They argued that a model-schools project be initiated in selected municipalities and that such projects be continued for at least five years. Further, they recommended that these projects be guided by a national committee appointed by parliament and emphasized the need to keep such projects independent both from the National Board of Education and from the pedagogical institutions at the universities and teachers' colleges. Kilborn and Lundgren viewed this suggested programme as a means of accumulating experience from increased research and evaluation. They recommended that as this experience accumulates over time the schools should be given more freedom to design staffing, grouping, and scheduling patterns to fit their local needs.

As a consequence of the evaluation study, the SIA committee changed its legislative proposal. In its final statement - a document of almost 1000 pages - it recommended that the government move toward a greater degree of decentralization than was tried out in the experiments. It stressed again that no ability grouping should be permitted, and that a continuous evaluation be done of every school, every school district, and of the national system as a whole. Also, as a result of the Kilborn-Lundgren evaluation study they recommended that this new reform be implemented in successive stages and that experience acquired in the initial stages be used in later stages to modify the other major aspects of the educational system.

The SIA legislative proposal was met by tremendous public interest. Earlier educational reform in Sweden had not been so widely debated in public. The evaluation of the experiments was itself widely discussed. The government prolonged the period before the legislative proposal was presented to parliament in order to get all opinions on the proposal. In the debate the evaluation report was discussed on television, on radio, and in a series of articles in the newspapers. All of the various teacher unions used it as study material. Most parent unions studied it also. The report has since been condensed and published as a pocket book.

Basically, the efforts of Kilborn and Lundgren were an attempt to document more completely and more accurately what had happened in each of the experimental schools and the consequences and implications of the reallocation of teaching resources. It was a post-mortem study, an evaluation in retrospect. Faced with a critical need for more information and with an imminent deadline, the parliamentary SIA committee solicited their help. Within a few weeks the plan of study was negotiated and within a few months the evaluation was completed, the report submitted, and the committee's legislative proposal redrafted.

The full story is yet to be told - in fact, the story may have only begun. What can be said is that a national debate over school policies has been aroused by an experiment and a legislative proposal. Further, it can be said that what linked the experiment and the proposal together in a most essential way was a modest evaluation study, a brief and retrospective analysis of the modified organization and teaching in the ten experimental schools.

REFERENCES/SOURCES

KILBORN, Wilko and Lundgren, Ulf P. Experiments with freer uses of resources. Göteborg, Sweden: Pedagogiska Institutionen, 1974.

LUNDGREN, Ulf P. Personal correspondence with Robert Stake. September 1974; January 1975.

7. ASSESSMENT OF PROGRAMMES OF EDUCATIONAL LABORATORIES AND RESEARCH & DEVELOPMENT CENTERS IN THE UNITED STATES, 1972

The Labs and Centers Program Evaluation has been compared to the Thirty Years War "...complete with a cast of thousands". The purpose of this case study is to provide a description of the negotiations, commissioning, and carrying out of this large-scale operation, focusing on those aspects of the 1972 assessment. To keep the case study brief we have omitted much of the story of the Laboratories Centers themselves. This is a case study of an evaluation - told here because the evaluation approach utilized and the problems encountered may be instructive to other individuals initiating a multi-institutional programme evaluation study.

In 1963, the United States Office of Education (OE) began establishing university-based "Research and Development Centers". The inquiry at each centre was expected to accelerate the understanding and improvement of educational practice with regard to a specific problem area.

In 1965, the Congress authorized OE to support the establishment of semi-independent, non-profit "Regional Laboratories" designed to make the results of innovation and experimentation in education readily available to the schools. These laboratories were to take research findings and from them develop educational products[1] that could be used in classrooms. They were to test and refine the materials and make them available to school systems.

Background: The Need for an Evaluation Study

In January, 1971, 13 university-based centres and 10 regional laboratories (hereafter referred to collectively as "Units")[2] were being federally supported through the Division of Educational Laboratories (DEL), a subsystem of the National Center for Educational

---

1) A "product" consisted of an "organized set of materials and procedures that may be used reliably to accomplish a socially desired function (Schutz, 1975)". Products included a variety of instructional aids ranging from books, transparencies, etc., to special teaching techniques designed to accompany materials.

2) Units were large-scale operations with their own physical plants. Some units had staffs of over 200 persons. They often were referred to as institutions.

Research and Development (NCERD), within the U.S. Office of Education (OE). Two or three nits established by other legislation had also been transferred to DEL jurisdiction. The quality of work being done at the Units had been criticized in some high government offices - in OE and the Congress, for example. In some quarters it was felt that the Unit budgets were not justified.

Prior to 1971, nine other Units had existed but had been closed. Those termination decisions were criticized. The then current DMI[1] funding policy seemed to be promoting destructive competition and attrition among the Units. The survival of the most promising Units had required a "midcourse termination" of other, less promising ones. New programmes were being funded before programmes already under way had been fully supported. The attrition had led to an insecurity and lowering of morale among the Units that remained. It was believed that this trend had to be reversed to strengthen the Labs and Centers Program.

Harry Silberman assumed the directorship of NCERD late in 1970. He was one of a number of individuals charged with preparation of a design for the proposed National Institute of Education (NIE). One anticipated NIE responsibility would be the annual allocation of funds to research and development efforts. Later, as part of the planning for the new institute, Silberman called for a thorough examination of previous funding policies, especially the bases on which funds had been withdrawn from some Units and not others.

Silberman knew that earlier evaluation efforts had been criticized on several grounds. For example, the most common evaluation procedure which had been used, a site visit procedure, was seen as lacking reliability. For each site visit a new team of researchers and experts was assembled, but given little background briefing. Also, inappropriate evaluative comparisons had been made between incomparable programmes. Site visiting evaluators usually reviewed <u>all</u> programmes at <u>one</u> Unit, comparing to some extent all programmes at that Unit with each other, in spite of the dissimilarities among them. Earlier assessment designs made little provision for simultaneous evaluation of similar programmes (e.g., <u>all</u> elementary mathematics programmes) <u>across</u> Units. (Other criticisms are detailed later in this case study).

It was believed that new evaluation procedures specially designed to overcome the shortcomings of previous methods were needed. Hopefully these procedures would enable decision-makers to determine which

(1) In March 1971, the Division of Manpower and Institutions (DMI) was created out of a reconstituted DEL. The labs and centres thus came under the jurisdiction of DMI.

Units should continue to receive programme funds. Silberman wished to design an evaluation and support policy which would transfer to NIE's jurisdiction only those Units whose programmes were deemed the very best. Reduced appropriations, new priorities, and inflation necessitated the cutbacks in support for existing programmes.

In June, 1971, DMI proposed a new support policy designed to reverse the attrition trend and to provide initiative for new programmes. The new policy (Frye, 1971) outlined a "maturity model", a system for allocation of funds to Units. Units categorized as not yet mature would be given institutional support (non-programmatic operating funds) from DMI. However, institutional support for new institutions was to be provided only during the Unit's early development. Mature institutions, according to the proposal, would be largely dependent upon funds secured by the Unit through programme contracts, from DMI and from other government and non-government sources.

W. Ray Rackley of DMI was asked to devise an evaluation system. He gathered opinions on perceived strengths and weaknesses of the earlier evaluations during consultation with NCERD staff, Unit personnel, and site visiting evaluators. Rackley had inquired and received advice on an evaluation plan from Daniel Stufflebeam, whose work in programme evaluation had gained OE's attention. Stufflebeam's system for generating an evaluation plan included the following steps:

1) a background study of existing evaluation procedures used in Unit evaluations;

2) specifications for the design of the new evaluation system;

3) alternative evaluation designs submitted by advocate teams;

4) the final evaluation design selected or synthesized from the alternative designs.

Stufflebeam, Director of the Evaluation Center at Ohio State University, wrote a working statement of this plan and in June, 1971, submitted it to OE. Later, it was funded but with substantial modifications. The proposal had outlined a 6 1/2 month schedule for completing the four steps. The NCERD personnel said that results were needed much sooner. The schedule for implementing Stufflebeam's plan was miniaturized - to two months - June 28 to August 28.

Stufflebeam's Advocate Team approach, for generating alternative evaluation designs, was aimed at encouraging the consideration of diverse philosophies of evaluation, each perhaps to be presented by national leaders in evaluation methodology. Two advocate teams were named, one chaired by Stufflebeam and one by Michael Scriven, both of whom had made substantial contributions to evaluation theory and practice.

Stufflebeam's proposal for generation of an evaluation system provided for a Convergence Team to convene after the Advocate Teams had each proposed an evaluation strategy. The Convergence Team had to operationalize either the selected strategy or a strategy combining elements from both proposals. It was to include Unit personnel and others who would eventually be responsible for implementing the system. However, lack of time precluded the use of such a Convergence Team.

Designing the Alternative Evaluation Systems

Scriven and Stufflebeam each selected three Advocate Team members during the first part of July, 1971. Since a Convergence Team was not used, DMI wished to have the Units' interests represented in some other way. Two representatives from the Units were named to serve as consultants for each team, selected by the team chairman from a list developed by DMI. Thus, each Advocate Team consisted of a chairman, three independent team members, and two consultants from the Units.

On July 24, 1971, team members and consultants met in Washington for an orientation session. Background data and instructions were provided. Before this session, Advocate Team members had been asked to identify documents needed to assist them in developing evaluation strategies. Diane Reinhard of Ohio State University had compiled questions that Advocate Team members wished to have answered at the session and had forwarded them to relevant persons in NCERD. At the orientation session, team members were given criteria for the evaluation procedures they were to develop. It was emphasized that the evaluation was to focus on the quality of the R & D programme or programmes being conducted at the Units and was not intended to examine institutional management capability. Each Advocate Team then met separately to discuss priorities and logistics and to talk with OE personnel.

In August, 1971, transcripts of the orientation session were sent to Advocate Team members and their consultants prior to their writing sessions. The Advocate Team chaired by Stufflebeam met August 5-9 in Phoenix, Arizona. The Advocate Team chaired by Scriven met in Santa Barbara, California, August 18-21. The one plan was drafted in Phoenix and the other was drafted by Scriven on the basis of the discussions at Santa Barbara.

The Final Evaluation Design

Copies of both Advocate Team plans were submitted to OE on September 1, 1971. Rackley, Program Monitor for the project, circulated the two plans and asked for critiques. Written responses were summarized. A panel of four independent judges was invited to assist NCERD decision-makers in their final choice of an evaluation strategy. On September 27, 1971, NCERD personnel, the Invited Judges Panel, and Stufflebeam and Scriven met at a morning session. In the afternoon, the judges met with the NCERD personnel to continue their discussions of the two designs.

The judges reportedly objected to the "overload of specific detail" in the Stufflebeam team plan - criteria and procedures for rank-ordering products, summing points, etc. The Stufflebeam strategy contained what one judge called "too many unconvincing numbers". The Scriven plan - described later - appeared to be more personalistic, a feature of the site visit procedure said to be objectionable. Neither plan, therefore, was entirely suitable. Silberman indicated a preference for the Scriven plan. Each had its supporters. The strategy developed by the team chaired by Scriven was named the winner, and, as modified with features from the Stufflebeam plan, was announced by Silberman the next day as the procedure for the 1972 Assessment of Lab and Center Programs.

The Scriven Design

The Scriven design called for two tiers of evaluation panels: a single Master Panel on top to which several Specialist Panels would report. Master Panel members would be trained to assume responsibility for decision-making in these circumstances. Master Panelists would then collaborate in the selection of Specialist Panel members. After training for their evaluation duties, each Specialist Panel would operate as a team of experts in the various disciplines or technologies. Each panel would submit its findings to the Master Panel for review. The Specialist Panels would be the fact-finding arms of the Master Panel, operating under its direction to examine matters that the Master Panel judged to need attention. The Master Panel would look across the board at all of the Specialist Panels' programme reports but would not be involved in any first-hand data collection. The Master Panel was to report directly to DMI and to the U.S. Commissioner of Education.

The Scriven plan, with contributions from the Stufflebeam proposal, prescribed the use of specified criteria in making programme assessments. However, the plan did not specify procedures, methods or data sources to be used by the panelists. The validity of the assessment system, as conceived in the Scriven plan, depended primarily

on the expertise of panelists as reviewers, observers, and judges. Heavy reliance was placed on the prior experience of panel members. The selection of a capable Master Panel was to be the first step in carrying out the assessment process.

Operationalizing Scriven's Design

Although not prescribed by the Scriven plan, the total set of programmes to be evaluated was divided into clusters. Personnel from DRDR (Division of Research and Development Resources, formerly DMI), NCERD,[1] and from the Units suggested ways of categorizing the programme. Formal classification of the programmes was new and of course several of the boundaries were quite arbitrary. The Units themselves decided their programmes would be best specified. Some Units described their entire work as a single programme giving it one title. Eventually seven clusters were defined. Each cluster contained a set of programmes, roughly content-related, which were to be evaluated relative to one another. Each programme was assigned to one of the seven clusters. Thus, programmes would be compared across Units rather than within. This feature of the implemented design was made in response to the criticism of previous evaluations which made comparisons among programmes that, in many ways, were not comparable.

The programmes in each cluster were content-analyzed and labelled as follows:

A. Theory and Knowledge Building for Organizational Change
B. Planning, Management, and Evaluation Systems
C. Instructional Personnel Development
D. Curricular Programmes
E. Culturally-Targeted Curricular Programmes
F. Home/School Intervention in the Early Years
G. Career Education

Persons with various content-area expertise were identified as specialists to judge the programmes in each cluster. About 80 experts accepted McVity's invitation to serve on the seven Specialist Panels. Orientation of panelists occurred at a 3-day meeting in April, 1972, in Washington, D.C.

The Specialist Panels then existed in the absence of the Master Panel - although, according to the Scriven plan, the Master Panel was to have created them. Commissioner of Education, Sidney Marland, indicated a preference for postponing the Master Panel phase of the evaluation until NIE was formally in existence, leaving options for final decisions on programmes open. The exact date of enactment of

---

1) The assignment of programmes to clusters occurred under the direction of Richard McVity, who had been Deputy Associate Commissioner for Research under Silberman. In December, 1971, McVity became Acting Director of NCERD following Silberman's departure to an NIE post.

legislation authorizing NIE was, at that time, still in doubt and Marland felt it was desirable to proceed with the Specialist Panel phase rather than delay the evaluation any further. The statements of purpose and a specification of the uses of the evaluation information - which the Master Panel was to compose as first steps in the evaluation - were therefore not available to the Specialists, and, in fact, were never developed. However, panelists were given a general orientation by NCERD as to the nature and purpose of their evaluation duties.

The Specialist Panel members were asked to take the general criteria developed by Scriven, McVity, and others and the oral briefings from the April meeting and to develop by themselves the specific criteria for programme evaluation in their cluster. It is important to recall that unlike earlier evaluations, this was an evaluation of programmes being conducted at individual Units, and not an evaluation of the Units themselves. Thus, while a Specialist Panel might recommend that a programme be phased out (or given extended funding), it was not in a position to recommend that a Unit be phased out (or given extended funding). A Unit with four programmes, for example, could conceivably have three of its programmes phased out, or even all four, and still remain as a Unit.

Prior to the formation of the Specialist Panels, Unit personnel had been required to prepare a document containing (1) an historical account of each established programme and (2) future development plans projected for five years, for all programmes, including those just being proposed and those already established. These documents, Basic Program Plans (BPP), were to be written according to guidelines established by DMI.

Specialist Panel members were provided with a BPP for each programme in their cluster. Products were collected in OE for panelist review. Face-to-face meetings with programme personnel were more the exception than the rule. At Santa Barbara, the Scriven Advocate Team had anticipated very direct contact between Specialist Panelists and Programme personnel and very direct observation of activities. The costs involved and the risks of co-option seemed so high, however, that the Specialists were told to examine documents only. Since each panelist had his own copy of the BPP and since for most programmes other documents were shared by panelists, even though it was a "proposal", the BPP became the major data source in the evaluation. For some programmes the BPP was supplemented with over 100 documents and products.

The methods of this review contrasted sharply with those of earlier evaluations of units or programmes. Site visitation had been the established procedure for evaluating programmes and products at Units. As an evaluation method, the site visit had been criticized both by personnel at the Units and by the site visitors. Often the qualifications of the visitors to judge programmes and products had been questioned. The brevity of the typical site visit also had been criticized. The concensus among Unit personnel was that a one or two day visit was an inadequate basis for making accurate and/or reliable assessments of Units. Furthermore, the site visits rarely had produced any useful formative evaluation information for the Units. Many Unit personnel had come to regard site visits as disruptive intrusions. In addition, there was the general concern that the personnel interchange, and sometimes conviviality, worked toward an unduly favorable judgment of the Unit. The recognition of these drawbacks was reflected in a decision to avoid the use of the site visit in the 1972 review, if possible.

Since the panelists would not obtain evaluation data by visiting the sites, they relied on programme documents, especially the BPPs, for evidence on which to base their judgments. Panelists were exhorted to insist on good evidence. An important shift in the burden of proof appeared to take place. In past evaluations, a Unit's effectiveness was more or less assumed until a site visit or product review revealed evidence of shortcoming. During this review, the Units had to provide documentary evidence of their efforts and their effectiveness. Unless a Unit could provide positive documentation of the efficacy of its programmes, it was assumed (under the Scriven plan as modified) that the Unit was not carrying out its required duties. In the 1972 assessment, funding was to depend on the Unit's documentation of the quality of its programmes.

It has been intended that each member of a panel would independently review and rate all BPPs within their cluster. This procedure also represented an important departure from previous efforts. Because the panel members were expert in more or less similar fields, each panel member was expected to contribute substantially to the recommendations for a programme. As the assessment progressed, however, it became apparent that for each basic question about a programme, one or two of the panel members were truly expert and the rest were not.(1) That fact, combined with the very large number of BPPs to be read, produced in some panels a division of labour.

---

1) Perhaps the most varied was Panel D (Curricular Programmes), the experts were an evaluation methodologist, an elementary school curriculist, a specialist in basic skills instruction, a reading researcher, a mathematician, an instructional researcher, a specialist in humanistic education, an artist, a music curriculist, and two specialists in data systems.

Individual panel members or subsets of the panel might have been assigned one or more of the BPPs to read, review, and rate in depth and then to report back to the whole panel the results of their review. It appeared, then, that only the "expert" reviewers rated a BPP independently. The other panelists arrived at their final ratings after the presentation by the in-depth review team. It is true, however, even with this division of labour that all of the panelists were expected to take responsibility for recommendations about all of the BPPs in their cluster. Most did, it is believed, looking particularly at those aspects of the programme that fell within their areas of expertise.

A Master Panel was eventually formed in June, 1972, shortly after the orientation of specialist Panels. It consisted of approximately fifteen members, all with affiliation outside OE. The Master Panel met for the first time in July, 1972, with Scriven as Chairman. Two or three subsequent meetings were held during the summer of the same year. Members discussed their responsibilities as Master Panelists and how they would deal with information from the Specialist Panels.

Specialist Panel reports were due in Washington on July 15, 1972. They were forwarded to the Master Panel. Each Specialist Panel Chairman made a brief oral report to the Master Panel. The Master Panel found themselves inundated with information and bound by time constraints, making impossible in-depth treatment of any single programme or Unit's programmes. In an effort to help the Master Panelists in their job of assimilating the evaluation data, a Synthesis Panel was formed to organize the Specialist Panels' reports into some manageable form. The Synthesis Panel attempted to summarize all Specialist Panel data and thus help bridge the gap between Master and Specialist Panels. The volume of the data and the lack of time available for dealing with it made the Synthesis Panel's task almost impossible. Synthesis Panelists attempted to reduce the complexity of the Specialist Panels' findings by rating all programmes on common criteria. Both the validity and justification for this data reduction were questioned. The Synthesis Panel laboured long but failed to simplify or objectify the tasks of the Master Panel or the then participating NIE representatives.

After completing their review, Master Panel members recommended one of four funding options for each programme: (1) continued funding for 3 years; (2) continued funding for 2 years; (3) continued funding for 1 year; or (4) phasing out of the programme. Recommendations of the Master Panel were forwarded to NIE Director, Thomas Glennan. As a result of the assessment, the funds for several programmes were terminated, in effect eliminating 5 of the 23 Units. Other programmes were given three-year contracts with increased funding. Total funding for all programmes was cut approximately fifteen per cent, as anticipated, because of reduced funding available to OE.

The directors of the Units were not pleased with the results of the 1972 assessment. In a list of objections they noted in particular that the assessment criteria were not securely based on the objectives originally stated in programme contracts. Objections from NIE personnel - most of whom were not involved even when the evaluation began - were partially in the opposite direction; that too little attention had been given to contemporary (1972) national needs for research and development.

The work of the panelists was individualistic, both calling for direction and resisting it. The evaluation had been an expensive experiment, resulting in severe cuts for some programmes. No one was highly pleased. Perhaps the best that can be said about it is that a conscientious, hard-working group of administrators and researchers - in panels, in OE, and in the Units - had taken a long and hard look at a massive R and D enterprise, and had used those observations to make some hard decisions. The Scriven plan was designed to be repeated over the years, modified by each experience. NIE did not use it a second time.

REFERENCES/SOURCES

COUNCIL for Educational Development and Research (CEDaR) Evaluation Committee. A review of the NIE evaluation of research and development centers and educational laboratories. May 1975.

DIAZ, Frank E. Interim process review of the 1972 spring assessment. Draft, undated.

FRYE, Charles. DMI institutional support and evaluation policy. Draft, June 3, 1971.

GORDON, Susan (Shana), Assistant to Acting Director of NCERD. Telephone interview, November, 1974.

MASON, Ward S. A proposal for a new institutional support policy. April 14, 1971.

MASON, Ward S. A proposed evaluation system for DMI. April 23, 1971.

MASON, Ward S. A critique of the right of passage review of laboratory and center program prepared for the mi-courses assessment. May 16, 1972.

MASON, Ward S. Issues related to the transfer of the R & D center and educational laboratory program to the national institute of education. July 17, 1972.

MASON, Ward S., Assistant Director of Research and Development Resources. Telephone interview, November 1974 and personal correspondence, January, 1975.

MCVITY, Richard. Guidance regarding NCERD fiscal year 1972 program assessment. March 29, 1972.

MCVITY, Richard. Next steps in the assessment process. April 24, 1972.

MCVITY, Richard, Acting Director of NCERD, December 1971 - August 1972. Telephone interview, 1974.

NCERD-DMI. A status report on the national program of educational laboratories, R & D centers and related R & D institutions administered by the division of manpower and institutions. April 12, 1971.

NCERD-DMI. (Work statement) Design of an evaluation system for the division of manpower and institutions. April 30, 1971.

RACKLEY, W. Ray. Charge to the advocate teams. July 9, 1971.

RACKLEY, W. Ray. Major steps in the design of the DMI evaluation system. June 29, 1971.

RACKLEY, W. Ray, DMI Chairman of Evaluation Task Force. Telephone interview, November 1974.

REINHARD, Diane L. Final Report: Design of a planning and assessment system for the division of manpower and institutions. August 31, 1971.

REINHARD, Diane L. Methodology development for input evaluation using advocate and design teams. Doctoral dissertation, Ohio State University, 1972.

REINHARD, Diane L., Ohio State University Evaluation Center staff. Telephone interview, November 1974.

SCHUTZ, Richard, Unit Consultant to Scriven Advocate Team. Personal correspondence, 1975.

SCRIVEN, Michael, Glass, Gene V. Hively, Wells, and Stake, Robert E. An evaluation system for regional labs and R & D centers: Advocate team report. August 31, 1971.

SCRIVEN, Michael, Chairman of Advocate Team, Professor of Philosophy. Response to questionnaire, 1974.

SILBERMAN, Harry, Director of NCERD, December 1970-December 1971. Telephone interview, November 1974.

STAKE, Robert, Member of Advocate Team chaired by Scriven, Professor of Educational Psychology. Personal interview, 1974.

STUFFLEBEAM, Daniel L. Proposal: Design of a planning and assessment system for the division of manpower and institutions. June 18, 1971.

STUFFLEBEAM, Daniel, Chairman of Advocate Team, Director of Evaluation Center - Ohio State University. Telephone interview, November 1974.

STUFFLEBEAM, Daniel L., Brickell, Henry M., Guba, Egon G., and Michael, William B. Design for evaluating R & D institutions and programmes: Advocate team report. August 31, 1971.

TRANSCRIPT of the Proceedings, Advocate Team Orientation Session. Washington, D.C.: July 24, 1971.

WELSH, James. Labs and centers: Cold winds from Washington. Educational Researcher, December 1972, 14-17.

8. AN EXPERIMENTAL APPROACH TO CURRICULUM EVALUATION: THE BSCS POPULATION GENETICS FIELD TRIAL

The Biological Sciences Curriculum Study was one of a number of large-scale curriculum projects initiated during the late 1950s and early 1960s. In these projects practising scientists sought to give the critical concepts and procedures of their disciplines a more central role in high school curricula. Prominent characteristics of many of the course materials produced were an emphasis on the structural elements of the discipline in question, an intuitive grasp of these elements by students, and an ability to apply them to unfamiliar situations. Initial leadership for the reform projects came largely from outside the education establishment. University professors in the academic disciplines involved themselves directly in problems of teaching school subjects in a way unprecedented in American education (Goodlad, 1966, pp. 32-41).

The history and accomplishments of the BSCS reflect these characteristics. The initiative for the formation of the BSCS came from the American Institute of Biological Sciences, largely as a response to the ferment of interest in curricular problems among scientists and mathematicians generally in the post-Sputnik years, and a dissatisfaction with the quality of biology being taught in American high schools. Although a number of leading science educators, that is, persons associated with colleges of education, were actively involved from its inception, professional biologists were prominent in the leadership of the project. It was thought especially important that high school teachers should be brought into the work, and they were not only consulted but served on the staff and the Steering Committee of the BSCS as well. Doing science, rather than learning <u>about</u> it, was considered important, and there was a premium on flexibility in the curriculum - three versions of the text with differing approaches to the same concepts were prepared, plus an array of open-ended laboratory exercises for the teachers and students (A. Grobman, 1969, Ch. 2).

Professional educational psychologists and learning theorists played a relatively insignificant role in the development of the BSCS curriculum,[1] and the procedures which were followed were something of a departure from those accepted at that time by most curriculum theorists. The difference is perhaps best suggested by Arnold Grobman, first Director of the BSCS, on reflection upon an early discussion of the Steering Committee:

> Persons with a background in Colleges of Education favored the construction of fairly full statements of goals and behavioral objectives and quite detailed outlines of content. The biologists without verbalizing it at the time, appeared to believe that the best way to describe the goals and objectives they had in mind would be to prepare a set of student materials, that is, textbooks and laboratory guides...such materials would indicate the goals more clearly than any abbreviated statement about them possibly could (A. Grobman, 1969, p. 75).

In the summer of 1960, after the general approach and the major conceptual themes of the curriculum had been settled, a team of writers was assembled to produce the first experimental versions of the three texts. The team consisted of equal numbers of specially selected high school biology teachers and university biologists. Writers had been paired (one teacher, one biologist) on the basis of expressed interest in a particular topic and version of the text, and began collaboration well in advance of the six-week conference. Even after the writing conference was underway, it was left up to the individual writers to decide how best to develop their particular topic and to present it to students. When the draft of a section had been completed it was circulated for the attention and criticism of other writers. A small group of high school students was also on hand to furnish reactions and to try out laboratory exercises as they were prepared. The following incident is illustrative of the way in which decisions were handled as they rose.

> On one occasion a special meeting of writers was called to resolve a particularly difficult problem which had arisen. Many of the high school teachers had become convinced that the materials were being written at too high a level of difficulty for the average high school student. Some of the biologists felt that since the purpose of the BSCS was to change the nature of the biology being taught in the schools, it would be a mistake to judge what students could do entirely on the basis of experienced with traditional materials. The high School teachers were persuaded that revision should await the results of an actual tryout in the schools (A. Grobman, 1969, pp. 90-91). The materials were found later not to be too difficult.

In this fashion, with much internal criticism and rewriting, three complete textbook manuscripts, with accompanying laboratory exercises and teacher guides were produced in just six weeks of intense and sustained collaboration.

Each text went through an extended period of testing and revision before a version was prepared for commercial publication. During the first year teachers using the materials were organized into testing centres. Individual teachers kept a log of reactions to the materials each day. The log format was deliberately kept open-ended, since structured questions might have discouraged teachers from recording unanticipated information: teachers were simply asked to report "whatever they felt would be useful to the BSCS" (A. Grobman, 1969, p. 108). Each centre was visited at least

once during the year by a "feedback consultant" to gather information about school conditions which might add to the significance of teacher reactions.

> Reactions of the participating teachers to a particular exercise were quite consistent and it was often rather easy for the visiting consultant to determine the fundamental reasons for those inconsistencies...that did appear ...(A. Grobman, 1969, p. 111).

Information collected in this way seems to have been the primary basis for judgments regarding revision of the experimental materials during a second writing conference.

> The value of the testing centers cannot be overestimated. The feedback reports supplied by the teachers were invaluable... (They) provided the BSCS with significant information on the utility of the materials in normal classroom situations... a key element in the success of the BSCS program. (A. Grobman, p. 116).

In addition to evaluation of this comparatively informal type the BSCS also carried on a substantial amount of formal testing. Testing was most intensive in the two years following the first revision of the materials; its purpose was to assess the extent to which students were acquiring the knowledge and skills indicated in the objectives of the curriculum. The first year of intensive testing (1961-62) was the major effort toward this end; there was an effort to obtain a representative sample of potential users and a control group of students in conventional biology curricula (BSCS Newsletter Nos. 10 and 12). The tests employed were of several kinds, but the major ones were a standardized high school biology test in use prior to the development of the BSCS materials, a comprehensive final examination based specifically on BSCS materials, and a specially developed "content-free" measure of understanding of scientific principles and scientific reasoning. (The latter two were developed by a special BSCS committee and were thought to be suitable measures of its objectives).

Results of the testing of 52,000 students in over 500 schools showed the BSCS students performing better on the BSCS comprehensive exam than the controls, but the controls performing better on the test of conventional content. Surprisingly, however, the control group performed better than the BSCS group on the "content-free" test of scientific reasoning. Further, BSCS students who had undergone a special supplementary laboratory experience - a six-week-long, open-ended investigation of a particular set of interrelated concepts - did not do so well on the BSCS final as students who did not have this experience (Wallace, 1963, pp. 22-24; H. Grobman, 1963, pp.25-29).

In the second year of testing (1962-63) the BSCS group as a whole performed higher on a revised version of the "content-free" test than did the control group. There was no indication given as to

whether the group receiving the special lab experience also outperformed the control group. But the non-lab group scored higher on both the BSCS comprehensive final and the scientific reasoning test than did the group receiving the supplementary experience (BSCS Newsletter, No. 20, p. 8).

Given BSCS's objectives of understanding scientific principles and ability to apply them to unfamiliar situations, and the heavy emphasis on active, relatively unstructured inquiry by the student, these results were the opposite of what was expected. (Of course it could have been that the tests used did not really measure these objectives). The results seemed to be contradicted by a wide variety of non-test indicators: Teachers, students, school administrators, reviewers in the discipline and the professional societies, and classroom observers agreed that the BSCS curriculum provided students with an impressively valuable experience. And other science teachers and college biology teachers found BSCS students much better prepared than students who had conventional biology instruction (H. Grobman, 1963; A. Grobman, 1969, pp. 137-41). No doubt the counterintuitive and conflicting results partly explain the repeated references in the writings of those associated with the BSCS to the difficulty of assessing the higher cognitive learnings and long-range outcomes which they had set out to promote.(2) Assumptions regarding the nature of an effective testing programme of a number of agencies either directly or indirectly concerned with BSCS policies seem to have contributed to this concern.(3)

Early in her tenure (1960-65) as Director of Evaluation, staff member Hulda Grobman began consulting with individuals in colleges of education around the country on problems the project had encountered. One of those with whom she consulted was Professor J. Thomas Hastings of the University of Illinois. Subsequent to these consultations the BSCS invited Hastings to serve as chairman of its Evaluation Committee. In this position Hastings saw his primary responsibility to be that of assisting the BSCS to broaden and improve the kinds of things done in evaluating its curricula (Hastings, 1974). In evaluating a curriculum he felt it was not enough, for example, to say that students "had studied" a particular text or "were taught" a particular course. If measured outcomes were to be interpretable, it was also necessary to give a careful description of the actual teaching acts and classroom conditions under which they were obtained. Similarly, the tests developed by the BSCS previously had followed the conventional design for tests which would identify the best and poorest students. Items had been selected to discriminate among examinees, rather than to discern what an examinee knew about the particular content and skills to be learned. There

were good reasons why those tests could not properly be used for making judgments about the depth of student understanding or the effectiveness of instructional materials, and the BSCS staff understood those reasons better than the testing specialists did.[4] (Hastings, 1967, 1974).

As the curriculum reform movement of which the BSCS was a part was getting underway, the field of education was feeling the impact of another movement, that of programmed instruction - an approach which advocated the analysis of any body of knowledge to be taught into specific bits of information. The BSCS had explored the possibility of adopting programmed instruction for its purposes as early as 1962. It was judged to be undesirable for entire texts or laboratory exercises, but the possibility of programming particular concepts was left open (BSCS Newsletter No. 13, p. 4; No. 17, p.11). When Hastings was asked by the BSCS to react to this possibility, he saw an opportunity to demonstrate something of the expanded potential of evaluation at the same time. He proposed that Professor Richard Anderson - a colleague at Illinois and one of the more sophisticated and skillful developers of programmed materials - be commissioned to develop materials which would test programmed instruction's potential for the kinds of learning sought by BSCS.

Anderson, however, was not only a skillful programmer. He was also a deeply committed empiricist. He held the idea that empirical validation was an essential part of the curriculum development process itself, not something to be brought in after the programme has already been developed (Hastings, 1974; Anderson, 1974b). Hastings knew that if Anderson were asked to develop materials he would simultaneously do a thorough evaluation job. Further, in addition to believing that empirical validation was an integral part of curriculum development, Anderson had a different conception of testing than the more conventional one in operation at the BSCS. He stressed that in evaluating curricular materials or an instructional sequence it is the student's knowledge and understanding of the subject in some absolute and particular sense that is critical. This is a different purpose from getting the best possible measure of which students, relative to other students, are showing the best overall general achievement.

The BSCS asked Anderson to develop a package of instructional materials. Their interest was primarily in getting information about the effectiveness of programmed instruction for teaching conceptually difficult material. Anderson's responsibility was to develop materials which could be used to test programmed instruction for this purpose. "Population Genetics" was selected as the topic to be taught.

It was one which teachers using the BSCS materials had found particularly difficult, and it was one not easily dealt with in the laboratory setting preferred by the BSCS. The primary reason for evaluation usually cited by the BSCS personnel was to get information for future revisions. But neither that nor the specific methodology to be employed in evaluating the materials was a subject of discussion between the BSCS and Anderson. As William Mayer, Director of the BSCS at the time the Anderson study was commissioned (still Director in 1974) said: "We felt him to be exceptionally competent in this area" (Mayer, 1974a).

Anderson recalls his discussions with the BSCS as having been rather hurried, of an informal sort: "I was pretty much given carte blanche; they wanted me to 'do something interesting' with programmed materials" (Anderson, 1974b). While the effectiveness of programmed instruction was his immediate concern, there were other issues and other audiences as well.

> The evaluation issue was 'in the air' at the time the arrangements were being made. Hastings was their outside evaluation consultant, and he was trying to influence the kinds of evaluation they did. A number of people had been brought in to do innovative things. (Ibid.).

Anderson saw himself as part of this effort. He was interested not only in influencing evaluation procedures at the BSCS, but also those of the other major curriculum projects. "I was interested in showing them that programmed materials are more useful than they thought, but the curriculum validation issue was the most important one to me" (Ibid.).

And, indeed, Anderson's approach to curriculum development and evaluation was rather different from that which the BSCS had employed in its earlier work. In particular, he was much more willing to place trust in formal tests and test data at every stage of the process. "Curriculum decisions should be based on hard data and fine-grained analysis; we have to look at the consequences of experimental materials in terms of student performance" (Ibid.). In this he saw himself as an advocate of a more rational and systematic approach to curriculum making. "I had a hard time convincing people at the BSCS, particularly the high school biology teachers, that hard data mattered. They were much more comfortable with their intuitions" (Ibid.). For him, leaving decisions about appropriate content, organization, and strategies to individuals who had only unexplicated experience to guide them was simply inadequate.

> At every step in the development process, those who claim to have produced something better should be asked for the proof of their claims.... there are no acceptable grounds for claims about effectiveness of a lesson except for data which actually demonstrate effectiveness (Anderson, 1969, pp. 4-5 and 29).

The practical implications of this approach came out perhaps most vividly in the way Anderson went about developing the programmed materials:

> ...As a first step, objectives were defined. In this regard, the treatment of population genetics in the Biological Sciences Curriculum Study textbooks served as guides. Next, a trial version of a portion of the program was prepared. This segment was tried with a series of individual high school students with one of the authors of the program monitoring each student's performance. After every few students, revisions were made. The remaining segments of the program were developed in the same fashion. Eventually the entire program was tested with small groups of students. Once again revisions were made. Throughout the development of the program, a lengthy criterion-referenced test - consisting largely of open-ended, constructed response items, problems to be solved, concepts and principles to be defined and illustrated - was used. Pilot subjects spent nearly as much time completing the criterion test as they did receiving instruction. Generally, a segment of the program was judged satisfactory when all pilot subjects scored 90 percent or better on criterion test items keyed to that segment. (Ibid., p. 9).

In the actual field trial the programme was compared with population genetics instruction given by teachers using one of the regular BSCS texts (the "yellow" version). Subjects in the trial were approximately 750 biology students taught by nine teachers in two suburban high schools. Half of each teacher's classes received the programme and half received the instruction ordinarily given in the BSCS curriculum. Because one of the questions to be answered in the study was whether programmed materials could be used with success under normal classroom conditions, no guidelines were provided for their use. Teachers were simply advised to use their "best professional judgment" in deciding how to use the materials (Ibid. pp.10-11).

An examination of the kinds of data collected, and the manner in which it was collected, can be an important clue to the factors governing the conduct of an evaluation. Given Anderson's belief that hard data are crucial, it is not surprising that test scores were one kind of information collected. The achievement test used to compare the learning of the two groups in the study was developed by combining 24 items on population genetics taken from the standardized BSCS achievement test, and 20 specially constructed "criterion"-type items (see footnote 4) of the sort used in the development of the experimental materials for the field trial. Both professional biologists and high school biology teachers associated with the BSCS were asked to judge the appropriateness of these special test items. As has already been indicated, Anderson was convinced that the only kind of testing that is valid for evaluation purposes is testing against a measure of knowledge or understanding of whatever is to be taught.

Ultimately, he hoped to persuade the BSCS and other curriculum projects to utilize criterion-referenced testing. The reason for including both kinds of items in his test was twofold: so that it would be credible to BSCS, and to provide an opportunity to compare the two approaches to testing.

Test data were not the only data collected, however. In addition, teachers were asked to keep a fairly detailed log describing classroom activities - laboratories, exercises, reading assignments, etc. - for both programme and no-programme classes, and the amount of time devoted to each. In contrast to the open-ended log that the BSCS had asked teachers in its testing centres to keep, this log was structured. It indicated the kind of information the researchers wanted. A questionnaire was developed and administered to ascertain teachers' and students' attitudes toward the programmed instruction.

The materials were used in the two schools for two weeks. On the average, students who had access to the programmed materials gained about 4 1/2 items between pre- and post-test, whereas control students gained about 3. Each student answered 22 items on each occasion. The difference was statistically significant but not large enough to be impressive to most people.

According to the teacher logs there was wide variation in the instructions given to the half of the class having the programming added to the regular instruction. Some teachers made the programme available but did not require its completion. Some required it but gave no class time for it. Some required it and allowed class time. When student achievement was analyzed as a function of those different procedures it was found that the students required to complete the programme and had class time to do it answered about 3 1/2 more items correctly than the other.

To allay his fears that the tests might favour the experimental group by testing them only on material directly taught, Anderson categorized the test items on a similarity to instruction scale. Some items required application of principles in examples different from those used in the programmed materials; some required the application of reworded concepts, and some required students to make inferences, etc. On all but one of the categories, that of application of principles to very different examples, the experimental students outperformed the controls.

It is worth noting that the non-test data were included to explain differences in test scores. They were not used as observations in their own right and they were not used to help the BSCS writer or other reader understand that field situation.

One of the audiences of the reports was, of course, the BSCS. A number of informal progress reports were submitted during the period of the study. These were found at the BSCS to be "very informative".

When the test data were interpreted a lengthy report was submitted. It was followed by intensive discussions between Anderson and Mayer and others at the BSCS. Both the substantive findings and their implications, of course, were discussed at these meetings.

The most visible reporting was directed at the educational research community. With full encouragement from the BSCS, Anderson presented the results in a symposium in 1967 before the National Society for Programmed Instruction. In 1968, a report was presented at the most prestigious meeting of test specialists in the United States, the Invitational Conference on Testing Problems sponsored by ETS. Only a brief and incomplete report was made to the faculty and staff of the two participating schools, until a year later when reprints of the published proceedings were available.

The BSCS held, and still holds, the evaluation of the population genetics materials in high regard. This is shown in a number of ways, some of which bear directly on programmed instruction, and others more generally on curriculum development and evaluation. A two-page summary report of the findings was published in the BSCS Newsletter, which went out to all teachers using the BSCS curriculum. As Mayer put it, "We had felt that (Anderson) had done a first-class evaluation job and the results were impressive enough to be communicated to the widest possible audience". It was decided that the population genetics programme should be published, and that a series of programmes on three other topics should be developed as well. These additional programmes were never published, however. One explanation was that the first programme did not sell well. There was some feeling that the publisher had not sufficiently promoted it (Steele, 1974), but Mayer attributed the commercial failure of the programme to a "decline in market acceptance", brought on by the fact that so much of the programmed material produced elsewhere had been of poor quality (Mayer, 1974b).

As had been indicated, Anderson was interested in influencing the BSCS's development and evaluation policies. He was at least temporarily successful in doing so. The BSCS invited him to serve on a writing team working on the revision of a text which had been in use for a number of years, and he was the first educational psychologist to serve in this capacity (Leibherr, 1975).

As a member of this team he was able to persuade the BSCS to undertake a study similar in methodology, but much more extensive in scale, to the population genetics evaluation. In this study a lengthy unit achievement test of the criterion-referenced type was administered to a representative sample of students from 60 schools where the text had been in use. In Anderson's opinion, analysis of student performance on the items in this test showed a number of problems with the materials.

Not all members of the writing team had confidence in the revealed defects, however. It is always possible to quibble about the quality of the test items which supposedly reveal the defects. Changes were made, however, and these were at least partially based on the test results" (Anderson, 1974b).

In a later phase of this more extensive study, test scores were subjected to factor analysis. This analysis, carried out by a statistical consultant brought in from outside by the BSCS, proved to be very costly - in man-hours and computer time - and rather uninformative for the BSCS's purposes. The supervisor of the writing team, himself a high school biology teacher, expressed disappointment that the study "did not point out more areas needing revision". He went on to express the view that, "maybe our earlier intuitive feelings about what kinds of information can be taught to students were better than might be expected" (Leibherr, 1975). He remained hopeful, however, that better research would obviate the need for reliance on intuitive judgments. As a result of this experience the BSCS largely lost interest in doing this kind of close and careful analysis of test scores as a way of making judgments about curricular materials. Recent BSCS evaluation reports feature, in addition to test scores, a variety of information from a range of information gathering procedures, and a relatively informal style of analysis and interpretation (Steele, 1974).

In conclusion, perhaps the most noteworthy thing about this evaluation was the diversity of audiences, questions, and considerations involved. Formally there was only one audience: the BSCS; and one question: do programmed materials teach biological concepts effectively and with understanding? The investigator was given autonomy with respect to his methodology, and the evaluation was extremely well received by the funding agency. Yet programmed materials did not assume the importance in the BSCS programme that the evaluation would seem to have justified, and the primary goal of the evaluator, that of persuading the agency to adopt highly analytic procedures for curriculum development and evaluation, went unrealized.

FOOTNOTES

1) Arnold Grobman reported in his history of the project that one of the first problems confronted was that of the absence of the linear structure in biological science which was to be found in the other sciences. The problem of "characterization of its structure appropriate for secondary school students" was extremely difficult (A. Grobman, 1969, p. 65). When educational psychologists were asked which of the several ways of going about this task would be most effective, "The replies we received...were, essentially, that the answers were not known, at least in a way that would be of practical application to the problems facing the BSCS. The psychologists recommended that we try both ways, whenever we had polar considerations...and that only through testing with students could we determine which promised to be the most effective procedure for us to use" (Ibid., p. 67). The only general meeting of the writing team during the first summer writing conference of the BSCS was devoted to a presentation by a learning psychologist "who discussed the implications of his field for members of the Writing Conference. Although the discussion was interesting, the writers felt it had but limited applicability to the task at hand". (Ibid., p.89).

2) In his history Arnold Grobman said, "One of the great hurdles faced by the BSCS in developing a comprehensive testing program has been to design some effective measures for the inquiry aspect of the BSCS courses. It is difficult to prepare tests to cover factual information and the handling of such information at complex conceptual levels. It is even more challenging to determine through multiple-choice tests whether the BSCS courses are contributing to an improvement in student facility for inquiry and investigation." Commenting on the failure of "block" students - those receiving the supplementary laboratory experience - to perform as well as "non-block" students, Hulda Grobman was even more explicit. "The BSCS has not been entirely satisfied with the quantitative measures of the effect of use of the block to supplement a BSCS Biology version. Non-quantitative evaluation feedback indicates that the blocks are a useful teaching device for enriching BSCS Biology for those teachers who elect to use it.... Thus, although major goals have been achieved, it is recognized that some of the BSCS objectives are not susceptible to quantitative measurement." (H. Grobman, 1963, p. 29).

3) Near the end of the first year's trial (1960-61) of the experimental version of the materials the BSCS formed an Evaluation Committee, with Subcommittees for Test Construction, Laboratory Block Test Construction, and Test Analysis. According to Arnold Grobman, the purpose of this reorganization was, "to design tests and examinations that would more closely reflect the principles of the BSCS program than had been the case up till this time...and to place a greater share of the responsibility for the content of the examinations under the control of biologists and to restrict the testing company's role to that in which it had special competence". (A. Grobman, p. 128). The following spring dissatisfaction was expressed within the Steering Committee with the way the Educational Testing Service, the outside testing consultant, had handled the testing programme (Ibid.). The ETS administered and scored the tests that year. However, another firm - the Psychological Corporation - was contracted to process the data and also to

assist during the next year (1962-63) with development of the tests to accompany the commercial version of the texts. (BSCS Newsletter, No. 19, p. 11, and No. 15, p. 12).

Another major source of conflict over testing was with the College Entrance Examination Board, whose tests had by this time come to be a major factor in determining whether high school graduates gained admission to college. The BSCS was concerned that its students might be unfairly penalized by a test which did not adequately assess BSCS objectives. The CEEB took the position that its tests should be oriented to the curriculum in widest use. Studies they had carried out indicated that BSCS students were not being disadvantaged by this policy. (A. Grobman, pp. 133-5, BSCS Newsletter No. 13, p. 2). However, after examining CEEB tests, two BSCS committees, who were "aided by independent reviewers from the fields of biology and psychometrics" concluded that "the CEEB tests do not adequately reflect the kinds of learnings that the BSCS hopes will be achieved through one of the BSCS Versions... The CEEB Biology tests we have examined do not provide full opportunity for BSCS students to adequately demonstrate what they have learned" (BSCS Newsletter, No. 22, p. 11). Writing in 1968 Arnold Grobman thought this conflict would eventually be resolved as the BSCS curriculum became more widely used; it would then be incumbent on the CEEB to make its exams reflect BSCS biology. There remained the "residual danger", however, that "even if based largely on the factual content of the BSCS courses, the CEEB tests could continue to be fact-oriented and thus still fail to test for the inquiry component of the BSCS courses" (A. Grobman, p.135).

The nature and extent of evaluation research to be carried out on the curriculum seems to have been a source of some difficulty with the BSCS's funding agency, the National Science Foundation. It was due in large part to the expectation of NSF, for example, that a test of conventional biology content was used in the major evaluation of 1961-62 (Hulda Grobman, 1975) even though "BSCS (did) not accept the aims of conventional biology courses - and this... is what the conventional tests have measured..." (BSCS Newsletter, No. 12, p. 10). The testing centers, from which the informal qualitative data on the classroom use of the initial BSCS materials had been obtained, were discontinued in 1962-63 because the NSF felt their utility had come to an end with the try-out of the experimental texts. The BSCS saw the centers as an important component of the curriculum, and would have preferred for them to remain in operation. (A. Grobman, pp. 115-6). Given the difficulties the BSCS encountered in evaluating the success with which students were gaining an understanding of science as inquiry, it is understandable that they should have been interested in continuing and improving the assessment of this crucial aspect of their programme. The NSF, however, did not see the support of educational research (as different from curriculum development) as part of its responsibility, and as a result "the optimal time for action slipped by without adequate steps being taken in Washington to support our evaluation projects" (A. Grobman, pp. 140-2, pp. 201-2; see also statements by Bentley Glass and Arnold Grobman on this problem in BSCS Newsletter, No. 27).

Finally, the BSCS was disturbed by much of the research which had been done on the BSCS curriculum by graduate students and professors in colleges of education. Taken as a whole this research was inconclusive with respect to the realization of BSCS objectives - for each study tending to support its expectations another could be found which raised doubts (Lehman, 1967). The BSCS

viewed much of this research (though not all) as "uninspired" (A. Grobman, p. 133). Its spokesmen tended to fault the methodology rather than the basic orientation of the research (See Mayer, 1967). In most of this research investigators used formal instruments to measure in quantitative terms what are essentially qualitative outcomes.

4) Several years later, at the prodding of curriculum developers, testing specialists ended a long disregard of the validity of test-content and the power of tests to indicate what actually had been learned. For years they had worked to develop tests with predictive validity, tests that would discriminate among students concurrently or according to some future prediction. These norm-referenced tests worked better, they had found, if not curriculum-bound, i.e. if the tests did not relate closely to any particular curriculum. But the 1960's brought the need for demonstrating to distant audiences the effectiveness of materials used by large numbers of students. Criterion-referenced tests, tests closely bound to a particular course of study, were the answer, at least for some developers (See Glaser, 1963).

REFERENCES/SOURCES

ANDERSON, Richard C. The comparative field experiment: An illustration from high school biology. Proceedings of the 1968 Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1969, 3-30.

ANDERSON, Richard C. Response to questionnaire. August 1974a.

ANDERSON, Richard C. Personal interviews, September 2 and October 24, 1974b.

BSCS Newsletter. Annual report, part B: Evaluation of BSCS biology. 1962, No. 12, 9-11.

BSCS Newsletter. New committee developments. May 1962, No. 13, 3-4.

BSCS Newsletter. BSCS committee activities. November 1962, No.15, 12.

BSCS NewsLetter. About BSCS biology. March 1963, No. 17, 7-15.

BSCS Newsletter. Preliminary report on the 1962-63 evaluation. February 1964, No. 20, 8.

GOODLAD, John. The curriculum. In John Goodlad (Ed.), The Changing American School, 65th Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1966, 32-58.

GROBMAN, Arnold. The changing classroom: The role of the biological sciences curriculum study. New York: Doubleday and Co., 1969.

GROBMAN, Hulda. Rationale and framework of the BSCS evaluation program. BSCS Newsletter, September 1963, No. 19, 9-11.

GROBMAN, Hulda. Some comments on the evaluation program findings and their implications. BSCS Newsletter, September 1963, No. 19, 25-29.

GROBMAN, Hulda. Personal correspondence, February 17, 1975.

HASTINGS, J. Thomas. A note on evaluation. BSCS Newsletter, January 1967, No. 30, 2.

HASTINGS, J. Thomas. Personal interview, August 1974.

LEHMAN, David. Abstracts of recent research and development: A new dimension in the evaluation of BSCS. BSCS Newsletter, January 1967, No. 30, 21-25.

LEIBHERR, Harold. Personal correspondence, January 30, 1975.

MAYER, William. Evaluation for curriculum improvement. BSCS Newsletter, January 1967, No. 30, 1.

MAYER, William. Personal correspondence, October 7, 1974a.

MAYER, William. Telephone interview, October 24, 1974b.

STEELE, Joe M. Personal correspondence, September 18, 1974.

STEELE, Joe M. Design and revision, data collection and portrayal. Me and my Environment, Formative Evaluation Report Number 3. Boulder, Colorado: Biological Sciences Curriculum Study, September 1974.

WALLACE, Wimburn. The BSCS 1961-1962 evaluation program - A statistical report. BSCS Newsletter, September 1963, No. 19, 22-24.

9. CASE STUDY OF THE EVALUATION OF PROJECT FOLLOW THROUGH

Origin of the Follow Through Programme

Project Follow Through was an outgrowth of the social policy initiated by the Kennedy Administration in 1960, and subsequently extended by both the Johnson and Nixon Administrations. Its roots were part of the "War on Poverty", and the general efforts of those in Washington during the mid-sixties to cure social ills by attacking their suspected sources rather than their symptoms.

During this period of history, a persistent cause of unfortunate educational experiences for poor children was their difficult transition to school environments. In light of this problem, Head Start was funded through the Economic Opportunity Act of 1964, and began as a pre-school compensatory education programme in 1965. Its intent was to provide disadvantaged children with new kinds of learning experiences, a variety of social services, and a range of medical benefits in the hope of easing their entrance into elementary school and equalizing pre-school educational opportunities. It was thought that Follow Through would extend the Head Start opportunities through the third grade. President Lyndon Johnson proposed it in his State of the Union Message on January 10, 1967.

Preliminary Development of Follow Through

It was initially intended that Follow Through would be funded by the Office of Economic Opportunity as an offshoot of the Economic Opportunity Act of 1964, and administered by the U.S. Office of Education.

As stated, Follow Through was to be a service-oriented programme similar in focus (but not necessarily in operation) to Head Start. During the spring a pilot phase was planned in some 30 demonstration projects in schools across the country. Evaluation specialists were consulted and a group of educational researchers at the University of Pittsburgh commissioned to evaluate the 12 month pilot phase. Funding of projects in 1967 ran well behind schedule but demonstration sites were set up. There was fairly general agreement among the Follow Through Programme staff and the planning staffs of OEO and HEW, that since the original budget request of $ 120 million was reduced to $ 15 million it would not make sense to carry forward with Follow Through as a service programme. Instead, the idea emerged to recast it into an experimental-demonstration programme.

The winter-spring of 1967-68 was apparently the period during which the notion of "planned variation" was refined. Active recruitment of project sponsors for Follow Through began as early as December 1967. There were no sponsored programmes in operation prior to the fall of 1968. Nevertheless, the recruitment of sponsors and the process of pairing sponsors and project sites was underway throughout the spring of 1968.

Several factors were ostensibly responsible for this shift in programme focus. Perhaps the most salient was the desired generalizability of the evaluation evidence that would eventually justify the support of a broader and more permanent programme that emulated and extended Follow Through's operational design. This desire strongly influenced the Follow Through implementation and, most significantly, its later evaluation design.

In An Organizational Analysis of Project Follow Through prepared for the Department of Health, Education, and Welfare by researchers at Northwestern University (Kaulee et al., April 1973, pp. 27-28) the transformation in programme emphasis from service-oriented to experimental was clarified. First of all, many of the evaluations of Head Start described the programme as either marginal or failing. Since Follow Through was originally set up to be similar to Head Start, such evidence was distressing. Second, there was the impact of the Coleman Report (1966) which questioned the effectiveness of fiscal expenditure as a means of equalizing educational opportunity. Third, and perhaps most important, preliminary evidence gathered on the 40 pilot Follow Through projects (the piloting occurred in 1967 when the service orientation was still dominant), although incomplete and not highly reliable, suggested only limited success at those sites (Kaulee et al., April 1973, p. 186).

One final point in this regard was noted by Robert Egbert (1973), Follow Through's second director. He suggested that in the fall of 1967, before the enabling legislation was enacted, it became known that the Office of Economic Opportunity would probably receive substantially less money than had been expected, and that Follow Through, as a new programme, would receive little, if any, funding. The decision was therefore made and agreed to by the Office of Economic Opportunity, Department of Health, Education and Welfare, the Office of Education, and the Office of Management and Budget that Follow Through - for the time being - should be an experimental programme designed to produce information which would be useful "when" the programme was expanded to nationwide service proportions.

The formal implementation of the Follow Through as an experimental programme began in the fall of 1968. Many sponsors began working with local project staffs several months prior to the opening of schools. About 200 Follow Through classrooms were funded that year.

Evaluation Rationale

It is difficult to address the issue of evaluation without briefly discussing the role of policy analysis and planned variation in the context of the Follow Through Program. Richard Elmore, in a paper prepared at Harvard University (April 1974, p. 2), suggested that the major reason for undertaking planned variation experiments was to inform educational policy. The basic principle underlying the notion of planned variation is relatively simple: deliberately different programmes were to be set up in a number of natural settings, and their effects were to be studied over time, comparing one approach with other approaches and comparing programme children with control-group children. Each approach was to be developed by an offered under contract through an educational research and development group, called a programme sponsor. Each sponsor would work at several sites. After selection and invitation by the local community, the sponsor was responsible for designing, implementing, and modifying the experimental project. They were to act as "change agents", responsible to the local school and community for developing and operationalizing an educational model commensurate with local needs. There were about twenty of these sponsors in Follow Through. Their approaches ranged from the highly structured, academically oriented to very informal, self-consciously non-academic open classrooms.

Policy analysis is predicated on the proposition that it is both possible and preferable for public policy decisions to be based on a rational choice among discrete alternatives. A goal for compensatory education programmes such as Follow Through, is to improve the school performance of disadvantaged children. In order to attain this goal, the policy maker should review the evidence on the effects of several approaches, selecting the most effective and abandoning the less effective. Policy could then be developed that would encourage school systems to adopt programmes from among those determined to be effective. A planned variation experiment offers the possibility of longitudinal evaluation in addition to data on contemporary choices.

It may be useful to point out that without the experimental nature of planned variation Follow Through, as a service programme, probably would have been limited to an undisciplined evaluation of a conglomerate of activities. Shifting to the planned variation design meant it should be easier to distinguish between sponsored projects that would succeed and those that would fail. Thus the probability that people would obtain an optimum amount of useful information from the national funding was increased.

Recently, in conversation, Richard Elmore, a diligent reviewer of Follow Through, now at the University of Washington, commented that social programmes, like Follow Through, should not be evaluated on a success/failure basis, even at the local sponsor level. Rather, evaluation efforts should generate hypotheses about the programme that could be debated in councils of action in light of broader social issues. But the enabling legislation was interpreted by Office of Education personnel as requiring success/fail evaluation comparing sponsored programmes against one another. The politics and process of that effort will now be explored.

Evaluation Design

In the spring of 1968 bids for the evaluation of planned-variation Follow Through were publicly solicited. The Stanford Research Institute (SRI) was awarded the contract in June 1968, to begin immediately. SRI is a large independent research corporation headquartered in Palo Alto, California. SRI's major evaluation tasks for Follow Through were set up as:

1. Design of the entire evaluation effort. (This was to be accomplished cooperatively with and subject to the approval of the Office of Education.)
2. Development of selection of appropriate tests and measuring instruments. (This also was to be a joint responsibility with OE.)
3. Data collection and storage.
4. Data analysis and reporting.

Officials, educators, and politicians were hopeful but not sanguine about SRI's initiative. There were disagreements in Washington and at the project sites as to which direction the evaluation should go. Would each approach be fairly treated? Poor results from the Pittsburgh evaluation of the unsponsored projects during the 1967-1968 year were just becoming known. Controversy over evaluation efforts of other compensatory programmes such as Head Start challenged both the methodological integrity and substantive narrowness of social intervention evaluations. The new design of Follow Through had greater integrity; but could it be carried out and evaluated?

Evaluation Difficulties

From the outset SRI was aware of the complexity of the evaluation effort and the politics that might confound it. The design task was exceptionally complicated by the range of approaches undertaken by different sponsors. It soon became apparent that many sponsors were

unable or unwilling to answer the evaluator's requests to clearly articulate what they hoped to accomplish. Since SRI was obligated to work with sponsors and community groups in designing evaluation strategies, the probability of achieving consensus regarding the design was severely limited. Despite such constraints, SRI managed to provide both evaluation services and evaluation information to a variety of audiences. They engaged in the collection of data for both formative and summative evaluations of both short run (up to one year) and long run (up to five years) effects.

SRI built a pool of outcome test items, partly by asking sponsors for items suitable to its particular approach. Logistical problems caused postponement of the classroom testing until the late fall of 1968. Once begun, however, the data collection effort was kept on schedule. In a recent correspondence Philip Sorensen, currently the Director of the Department of Educational Research for SRI, pointed out that the 1968-1969 year was later officially designated as a developmental year (i.e., children entering the programme in the fall of 1969 were then to be the first evaluation group), but this definition was not adopted until the spring of 1970 - the definition was ex post facto. In retrospect, it is difficult to understand how sponsors, the Follow Through Program staff, or SRI could have expected to implement programme approaches and an efficient evaluation during that first year (1968-1969). Perhaps SRI would not have been awarded the evaluation contract if it had been more realistic. The importance of a planning and developmental year for new programmes and their evaluation seems to be generally accepted today, but the start-up problems were not so fully appreciated in 1968.

Nevertheless, other problems emerged. Repeatedly, during the 1968-1969 period, sponsors complained that too much emphasis was being placed upon cognitive achievement gains by the SRI testing instruments. Additionally, the first year witnessed an increase in both classrooms and pupils (from 203 to 1,512 classrooms and from 4,059 to 34,319 pupils). Such growth constituted a crisis for SRI. More testing supervisors had to be located, more tests administered, at more sites, and far more data analyzed. This kind of logistical problem forced SRI to reorganize its Follow Through operations in November of 1969. The main organizational change was to establish a home-based full time staff for field operations and to discontinue the supervision of data collection responsibilities by part-time persons (e.g., university faculty) hired locally or regionally. The transition actually took several years to complete.

In addition to the logistical problems within the project, SRI was overcommited by attempting at the same time to do an evaluation of Head Start. Sorensen (again in informal correspondence) revealed that SRI began the Head Start work with the expectation that they could manage both projects with an integrated staff under a single project management. This soon proved to be unfeasible, due in large measure to calendar differences between Head Start programmes and the schools - it was impossible to use the same field staff on both projects.

A review of the SRI evaluation efforts by the Department of Health, Education, and Welfare's Audit Agency (p. 8) raised serious questions about SRI's ability to perform both evaluations simultaneously. This questioning led to three reviews of contract performance during the summer of 1970 and the spring of 1971. It is somewhat unclear from the audit reports whether the problems were with the Head Start contract, the Follow Through contract or both. In each review, however, SRI was deemed to be doing a satisfactory job under prevailing circumstances. However, the reviews did reveal that the "joint-venture" arrangement between SRI and the Office of Education regarding the Follow Through evaluation was not working. The prime difficulty arose from delays in getting procedural approvals from Washington. It was also noted that SRI had been asked to respond to unrealistic deadlines. These problems notwithstanding, SRI was retained as the prime contractor for the 1970-71 school year.

During this troubled period when efforts were being made to clarify and redefine the roles of Follow Through agencies and contractors, there was the additional difficulty of adjusting to the administrative changes within the Office of Education. In Jure 1971, responsibility for the Follow Through evaluation was shifted from the Follow Through Program office in the Office of Education to the Office of Program Planning and Evaluation of the Bureau of Elementary and Secondary Education in the Office of Education. At the same time, Robert Egbert, after providing critical leadership as Director of Follow Through, left the Office of Education, having announced his intention to do so the previous November. The monitoring of the project changed radically and abruptly. There was an almost total turnover of staff within the Office of Education who were responsible for Follow Through.

In November 1971, Garry McDaniels came to the Office of Education from the University of Maryland. One of his immediate responsibilities was to determine how the Follow Through evaluation should proceed. McDaniels was instrumental in retrieving the Follow Through evaluation as a responsibility of the Follow Through Program office. This occurred officially in early December 1971. Earlier in the year SRI,

by agreement had its Follow Through activities limited to data collection. The Huron Institute received a contract to provide a summary review of SRI evaluation efforts, and also consult on further evaluation designs and sampling procedures. Coincidently, the decision was made to separate the data collection and analysis contracts beginning in July 1972. A Request-for-Proposals was issued early in 1972 for the overall analysis. SRI proposed, as did many other groups. The award was made to Abt Associates, a Cambridge, Massachusetts, group, in the spring of 1972. Thus by that spring McDaniels had engineered the following changes in the Follow Through evaluation effort (Krule et al., 1973, p. 185):

Task	Past Responsibility	As of Spring 1972
Instrument Construction	SRI/OE	OE
Battery Selection	SRI/OE	OE
Sample Selection	SRI/OE	Huron
Data Collection including classroom observation and processing	SRI	SRI
Analysis	SRI	Abt

The restructuring, however, did not end the controversy. Another crisis occurred in May 1972, regarding SRI's role in the Follow Through evaluation. A letter from Ralph Nader's Center for the Study of Responsive Law was sent to Elliot Richardson, then Secretary of the Department of Health, Education, and Welfare. The letter questioned SRI abilities to perform any part of the Follow Through evaluation claiming several instances of malfeasance. Richardson immediately replied and explained that the scope of SRI's involvement had already been reduced. He also initiated another review by the Department's Audit Agency. This review exonerated SRI from any major breach of contract. SRI personnel protested Nader's accusation as irresponsible in several important respects and not deserving of the investigation it provoked.

After restructuring of the evaluation effort, McDaniels left the Office of Education for a position in the National Institute of Education. There is evidence that this and similar shifts in Office of Education personnel will again change the focus of the Follow Through evaluation. Abt Associates' contract for data analysis expired in August 1974. Also, in August SRI delivered its most recent technical report entitled, Follow Through Classroom Observation Evaluation 1972-1973. Allegedly, new Requests-for-Proposals to continue the evaluation are forthcoming. Thus, an evaluation with a data base originating in 1968 could be subjected to yet another design, a new set of questions, and an entirely new audience.

Evaluation Lessons

It should be clear that a fair and comprehensive analysis of the Follow Through evaluation requires more information than appears here. For example, the methodological issues raised in evaluating broad-scale social interventions, like Follow Through, are the worthy subjects of analysis. Philip Sorensen summed it well in a paper at the American Educational Research Association Annual Meeting, 1973, when he stated that, "Follow Through should not be viewed as an experiment in the classical sense. From the evaluator's viewpoint, it is a quasi-experiment - or even a pseudo-experiment - in which there is literally no manipulative control of treatments. It's a social experiment on a fairly ambitious scale with all the blooming, buzzing confusion that characterizes efforts to be systematic in a natural setting" (Sorensen, 1973, P. 7).

Similarly, the evaluation itself could be analyzed from an economic or social policy perspective. The attempt here has been merely to highlight some of the major forces that appear to have influenced the evaluation of Follow Through. With that in mind, what follows is a listing of occurrences in government which constrain such evaluations:

Preconceived notions by the contracting agency as to the outcome of the evaluation.

The imposition of unrealistic deadlines by the contracting agency.

The potential delays caused by a "joint decision-making" or "joint-approval" arrangement between the contracting agency and the contractor.

Shifts in evaluation philosophy, central questions, or usage caused by shifts in personnel within the contracting agency.

Reductions in budget of the contracting agency.

Too many audiences influencing the design of the evaluation and reporting of findings by the contractor.

The threat of constant review by the contracting agency or their designated representatives.

The threat of review by "public interest" groups.

Each of these issues are worthy of continued exploration, particularly in the context of the Follow Through evaluation. Such occurrences have been obstacles for SRI. It is only a matter of speculation as to whether a valuable evaluation study would have been completed had the contracting agency been faultless in its part.

REFERENCES/SOURCES

EGBERT, Robert. Follow Through, Unpublished paper, 1973.

ELMORE, Richard. "Planned Variation Experiments as Policy Analysis", prepared for the Annual Meeting of the AERA, Chicago, April 1974.

HEW Audit Agency. Review of Head Start and Follow Through Evaluation Contracts with Stanford Research Institute.

KRULEE, Gilbert K., Hetzner, William A., and McHenry, Emmit J. An organizational Analysis of Project Follow Through. Evanston, Illinois: Northwestern University, 1973. Report submitted to U.S. Government, H.E.W. Contract OEG-D-70-4940 (286).

SORENSEN, Philip. "Conflicting Viewpoints that Affect Design, Analysis, Interpretations, and Reporting in the National Follow Through Evaluation", presented to the Annual Meeting of the AERA, New Orleans, 1973.

Office of Education. Request for Proposal: Follow Through. Proposal request for analysis of Follow Through data collected through the spring of 1972.

Stanford Research Institute, Summary of Plans for Longitudinal Evaluation of the National Follow Through Program, 1969-1970, September 1969, pp. 19.

James L. Wardrop, Analyzing, Interpreting, and Evaluating the Data from the National Evaluation of Follow Through. Proposal in response to the RFP to analyze FT data collected through the spring of 1972. Submitted to HEW/Office of Education by the Board of Trustees, University of Illinois, Urbana, Illinois.

Stanford Research Institute, Longitudinal Evaluation of Selected Features of the National Follow Through Program. Menlo Park, California: Stanford Research Institute, no date indicated, draft of evaluation covering data through 1970, pp. 215.

Gilbert K. Krulee, William A. Hetzner, and Emmitt J. McHenry, Program Evaluation: The Follow Through Experience. Evanston, Illinois: Northwestern University, 1973, pp. 105. Paper based on experience gained through involvement in Follow Through evaluation contract OEG-D-70-4940 (286).

GARRY L. McDaniels, The Current Status of Follow Through as a Social Experiment. Expanded version of a paper originally prepared for a conference on social experimentation sponsored by the Brookings Institution. Draft, 8-20-73, pp. 56.

Copies of correspondence obtained from Center for Study of Responsive Law. Items listed below:

Letter dated May 24, 1972, from Daniel Guttman (CSRL) and Barry Willner (CSRL) to Elliott Richardson (HEW).

Letter, date not readable, from Elliott Richardson (HEW) to Center for Study of Responsive Law.

Press Release, May 26, 1972, Center for Study of Responsive Law. Concerning May 24 letter to Elliott Richardson outlining charges against Stanford Research Institute.

Press Release, SRI, no visible date, concerning CSRL allegation.

Telephone conversations and correspondence concerning the evaluation(s) of the Follow Through Program with the following:

Marvin G. Kline, Abt Associates, September 1974.

Dr. Garry L. McDaniels, (HEW) National Institute of Education, September 1974.

Dr. Richard Marciano, Stanford Research Institute, September 1974.

Dr. William Krulee, Northwestern University, October 1974.

Mr. Richard Elmore, University of Washington, January 1975.

Dr. Philip Sorensen, Director Educational Research, SRI, January 1975.

10. CASE STUDY OF THE GERMAN ADAPTATION OF THE AMERICAN "SOCIAL SCIENCE LABORATORY UNITS"[1]

The following report summarizes the experience of a project group adapting American high school curriculum materials for West German schools. The materials were the Social Science Laboratory Units (SSLU)[2] developed at the University of Michigan by Ronald Lippitt, Robert Fox and Lucille Schaible.

On a visit to the United States in 1967 Hartmut von Hentig, a professor at Bielefeld University, became acquainted with work on the SSLU materials. In 1971 he introduced the Lippitt-Fox-Schaible ideas to the newly gathered faculty of pedagogy at Bielefeld. Those ideas were built into a project, a whole-staff project, and in fact the only specially funded project then available for faculty research participation. All members participated - in fact, participation was expected of them. They decided they would adapt the American materials for German use. Permission was obtained from the Michigan group. The project was funded by the Bundeszentrale für politische Bildung (Federal Centre for Political Education). Keeping in step with the democratic ethic of the SSLU, the 15 persons of the staff - including 3 students - decided to become co-equal authors. For business purposes, von Hentig and Professor Helmut Skowronek were named to represent the group, as in a contract with a publishing house. Von Hentig became known as the programme director, but he and the group understood that there would be equal responsibility and authority by all members of the team.

Each staff member worked on the adaptation of the units of SSLU. Some worked on several units. Five members served on the Editorial Committee, but all 15 participated in important editorial decisions.

---

1) This brief case study is based mainly on two preliminary evaluation reports by Wolfgang Neul, two items from Hartmut von Hentig, and personal conversations. Karin Sitte assisted in preparing the case study.

2) In West Germany the project is recognized by the letters SfK, for Sozialwissenschaften für Kinder.

It was felt that someone on the team should keep a careful record of the progress of the project for the benefit of the staff as well as those interested outsiders. Wolfgang Neul perhaps was the only one with both suitable qualifications (a master's design in psychology though no experience in educational evaluation) and interest. He agreed to take the formative-evaluation responsibility. Although it was not put into writing, it was expected that the evaluator's task was to design instruments and set up certain heuristic devices for the adaptation work. Neul also worked on developing one of the units. There was no distinct division, either by personnel or by function, between adaptation and evaluation. Evaluation was to be an integrated function within the whole adaptation process.

Neul was different in that he worked full-time on the project - the rest worked part-time. Their main job was research and instruction at the University. The 15 team members were subdivided into panel groups of two or three. A variable amount of time was spent by each panel on its unit. Enthusiasm and competence varied too. The units differed in length and emphasis. A single theme or "story line" did not run through the units, nor were there common criteria for the adaptation. Cooperation among the panels happened by chance or not at all.

Group sessions varied from daily to weekly meetings and at times only once per semester. The long intervals made precise goal definition almost impossible. The different value positions were reinforced by the lack of continuous dialogue.

The following evaluation report is based on the trial version of the adapted materials. (The final version was not ready for publication at the time of writing).

THE EVALUATION PROCEDURE

The evaluation took place in four stages:

I. The orientation phase (5 1/2 months)
II. The preliminary evaluation (8 months)
III. The planning of the final evaluation (6 months)
IV. The final evaluation (5 months)

I. THE ORIENTATION PHASE

During the orientation phase the 15 team members (3 professors, 9 research assistants, 3 students) made themselves familiar with the project materials. In this phase there was no particular emphasis given to evaluation, nor was any need or desire for it apparent. Finally it was agreed that an evaluation should take place for (a) innovation, (b) legitimation, and (c) consistency reasons.

a. It would be irresponsible not to evaluate the innovative effects of one of the first large social-science curricula in West Germany.

b. It had to be proved to team members as well as clients and audiences that the adaptation showed at least first signs of success.

c. If questionnaires and interviews were procedures recommended in the laboratory materials, it was felt that they could not reasonably be rejected as unsuitable for evaluation purposes.

The evaluator was not regarded as representing the interests of teachers and students to which the curriculum was addressing itself. Thus by misunderstanding and disregarding one of the main functions of the evaluator from the very beginning, the team deprived him of one kind of authority which was needed in later curriculum decisions.

II. THE PRELIMINARY EVALUATION

In the second phase the evaluation group set itself three tasks:

a. planning and carrying out a preliminary evaluation strategy;

b. developing, applying, and revising instruments to test the effects of the materials;

c. evaluating and feeding back data to the other team members.

There was a general agreement on this procedure because:

a. there was a need to find out whether this American curriculum would encounter comparable preconditions in German schools;

b. hypotheses had to be developed for the final evaluation;

c. methods and instruments had to be tested.

Through personal contacts three schools were selected (an elementary, a comprehensive, and a grammar school) in which the first materials were to be tested. A training course (weekend seminar) introduced the teachers to the project.

The materials consisted of an abbreviated version of unit 0[1] (for student motivation), unit 1 (introduction to certain strategies and methods), and unit 7 (application of strategies for decision-making).

Each (of eight) classes was to have 12 lessons. To evaluate the outcome, the following instruments were used:

1. Teacher questionnaire (handed out at the beginning of the lesson). It dealt with the subject-specific training of the social-science teachers and the time they would spend in lesson preparation.

2. Student questionnaire (handed out at the end of the lesson). It dealt with the interest the students took in the programme and their self-assessment.

3. Programme questionnaire which raised questions on each particular lesson.

4. Observation checklist for classroom observation.

5. Informal protocols on classroom observation aimed at identifying standard conditions.

6. Student interviews at the end of the project.

Due to the lack of time there was no achievement testing.

The outcomes of the pilot phase can be divided into direct and indirect results. The direct results concerned student and teacher reactions:

The reactions to the programme seemed to be mainly reactions to the unconventional teaching style. In schools where inquiry methods and group discussion were rare the SSLU units caused a considerable change of motivation and interaction. In schools where inquiry learning and group discussion were common the SSLU units added to the existing tedium.

The evaluators have come to the conclusion that the materials should contain a wide range of approaches which could be used in a specific situation. Situations should be identified to which specific instructional approaches are appropriate.

Teachers expressed a strong desire for feedback. Detached observation has become more and more impossible. Consequently, a closer interaction between curriculum developers and teachers should lead to an immediate influence on programme development. Evaluation would then be a permanent micro-evaluation. If there is a need for

---

1) Unit 0 was added in the German version.

generalizable and comparable results, the evaluators should identify those programme instructions which lead to misinterpretation or leave teachers faced with unexpected situations.

As to instruments, the programme questionnaire, the informal protocols, and the student questionnaire were so instructive that they were used in a modified form of the final evaluation. The observation checklist was less profitable since the observation instructions and the observation categories were unsatisfactory. Thus it was considerably altered in form and function for the final evaluation.

The indirect results concerned group intersections and procedural decisions:

> Since the preliminary evaluation put members under considerable time constraint, they were forced to concentrate their efforts. However fruitful that was, it led to a growing tension between the adaptation and the evaluation team.
>
> The information gain of the evaluation group was not shared by the other team members. A "jour fixe" for information exchange seemed to be an unattractive arrangement. The adaptation group became totally involved with the construction of the units. Thus the lack of communication between the two groups reached a point where the evaluators were almost ready to resign. However, agreement was reached that the evaluation group should try to do more to inform the adaptation group about their procedures and to design evaluation instruments which could be understood and accepted by the others. Questions which might have been important to some people but were of little consequence to the present audiences were not to be raised. The evaluators then tried to make the procedures less rigid and more probing in order to stimulate more of an appetite for information on the adaptors' part. They asked the adaptors to give a precise definition of their goals while constructing the units so that both teams could develop and examine the hypotheses that followed. This request was not fulfilled. An important chance for the achievement of an overall focus for the units was missed. On the part of the adaptation team there was no willingness to question their own results, and proposals for revision were rejected on the following grounds:
>
> a. the limited scope and ambiguity of data did not warrant a large-scale revision;
>
> b. revision of units should not be based on findings from a school situation where the proper conditions for learning have not yet been established.
>
> The project participants committed themselves only halfheartedly to the project, perhaps a result of the oversupply of professional persons in the education departments of West Germany. Most of the team members regarded the project as "on-the-job training" following the award of a bachelor's or master's degree in subject areas only loosely linked with educational science.

III. THE PLANNING OF THE FINAL EVALUATION

The major aims of this phase were:

a. selection of teachers who were willing to test the materials in the trial classes;

b. selection of new samples of schools and classes which met the requirements of the particular teaching strategies;

c. revision of the already-existing and development of new instruments.

Experience during the previous stages had led to the common agreement that all procedures which the evaluation team developed or suggested had to be considered and approved by the other team members. Although this caused long and enervating discussions, it at least increased problem-awareness among the members.

a. Teachers for trial classes were selected through advertisements in three West German educational periodicals and through the teacher-training seminars in North-Rhine-Westfalia. Since participation was voluntary, a bias favouring innovation was to be expected. Teacher-attitude data were gathered for correlation with student achievement.

b. As far as schools were concerned, comprehensive schools were much more interested than grammar schools. Sampling became difficult. An attempt was made to teach each unit in each of the selected schools and to each age group at least once. The plan below shows the distribution of units. Since the type of school had proved to be an important variable in the teaching of the programme, the pattern below seemed to be the most appropriate.

	Grade 5	Grade 6	Grade 7
grammar schools	units 2 + 7	units 3 + 4	units 5 + 6
elementary schools	units 3 + 6	units 2 + 5	units 4 + 7
comprehensive schools	units 4 + 5	units 6 + 7	units 2 + 3

The development of test instruments was hampered by the lack of sound information. At the time they ought to have been constructed, the development team was still working on the goals of the project and implementation plans. Therefore, instruments were developed to test the global effects of the project on the one hand and a few particular items on the other hand. In addition, an attempt was made to gather data on the different conditions that might influence the effects of the programme, such as school "climate", attitude and education of teachers, percentage of working-class children per class, and previous training of classes.

The observation scheme for the preliminary evaluation was changed and adjusted to reduce the observation time. For each, of the three observations, teachers were to be asked to give a short outline

of the intended instruction: Did they structure the lesson according to materials, course objectives, or previous experience of students? Subsequent discussions with observers would focus on discrepancies between intended and actual procedures and on the implications for further course planning.

Teachers would receive a schedule which contained the obligatory and the optional paragraphs of each unit. This schedule had become necessary because the time limit did not permit a complete treatment of the units. The units had become more extensive than originally planned (only certain deviations from the American original were permitted by contract).

The teacher questionnaire from the preliminary evaluation was standardized: recurring questions were separated as to those referring to the textbook and those referring to the students' exercise books. The questionnaire was to check into observation of the schedule and reasons for deviation from it. It also was to find out the time allocated to the various sections.

In order to measure the effects of the programme in the affective domain, student questionnaires were to be handed out at the beginning and end of the programme. They would gather data on attitude change. The semantic differential was to be used to measure the concepts of "school", "teacher", and "scientist". In order to assess individual achievement relative to the socio-economic status of the family, data on parents' occupations and their interest in the students' homework were also to be gathered.

Student achievement in the cognitive domain was to be tested by seven unit-specific tests. The tests were developed jointly by the evaluation and adaptation teams. Setting up the objectives and developing test items which matched the objectives for each lesson was a profound learning experience for both teams. When unit questionnaires and tests were completed, they were sent out for review by interested teachers and other experts, who presented written reactions or oral judgment at one of the weekend seminars.

IV. THE FINAL EVALUATION

This phase focused on three major activities:

a. the weekend seminars for teachers and development team;

b. the administration of the instruments (described in the previous section);

c. the critique of the evaluation strategies.

a. At three weekend seminars, teachers and development team discussed controversial issues concerning the project and its implementation. The first meeting was held before the programme started. The second took place after units 0 and 1 had been taught and the last at the end of the programme. Each of the 20 teachers who participated in the programme took part in at least one of the seminars. Experience with the first two seminars showed that participation increased if the invitation contained a structured agenda with emphasis on teacher-identified problems.

The first seminar served to clarify the relationship between the teachers and the development and evaluation teams. Discussion focused on the developers' intentions. The critique by the teachers was aimed less at the implementation of the programme than at its political legitimation. They expressed doubts about the feasibility of the overall aims of the project.

At the second seminar, criticism was based on actual teaching experience. It focused on the lack of

1. a "red thread" (consistent theme) running through the units;
2. easy readability of textbooks;
3. variation of learning situations;
4. step-by-step teacher aid on some units;
5. clarity of arrangement of topics.

The third seminar was marked by a strong need for the exchange of information: Did the programme cause any change in the students' behaviour? Did it launch any action on the students' part? Could it be shown that the students actually made use of the theoretical information conveyed to them? There was little evidence available, and disappointment was obvious.

The attitude of parents was reported to be generally very positive. A few objected to the questions on family background as a violation of privacy. A "guidebook for parents" might help to clarify matters and attain cooperation.

The attitude of teachers who did not participate in the project varied. Again immediate information seemed to be essential. There seemed to be an apprehension that the SSLU dealt with problems which so far had been ascribed to other subjects such as religious education.

b. Difficulties arising from the application of evaluation instruments (tests, questionnaires) were due to:

1. heterogeneity of classes as to test experience;
2. demand for legitimation of procedures;

3. amount of extra work for teachers;

4. lack of contact between development team and trial classes;

The completion of questionnaires seemed to highlight a principal dilemma of programme evaluation: In order to gain comparable data, questions had to be standardized. This was only possible if teaching strategies were more or less prescribed and followed. On the other hand, data of the teachers' individual choices and procedures were equally important. Thus the questionnaire had to provide questions for a rigid as well as a flexible treatment of materials. That made it voluminous and time-consuming. Furthermore, the standardized questions seemed to be inadequate at a time when teachers needed immediate feedback from the developers rather than summations from a set of generalized questions. In general, data feedback was much better when the contact between development team and schools was close. Data feedback decreased if questions and tasks remained the same throughout a particular length of time. It was increased, on the other hand, when questionnaires and tests had an "official" image (e.g., on printed computer cards).

c. The incompleteness of data made an interpretation and evaluation of the overall situation difficult. Criticism focused on the non-representativeness of data and on the lack of explicit hypotheses. At this stage it became obvious that the previous disregard of the evaluator's demands and offers had caused many of the present deficiencies. In the very last stage the development team was willing to accept help from the evaluator - and even demanded it. But the lack of underlying concepts of evaluation and curriculum adaptation made a different approach impossible at this time.

After a working period of two years the bulk of the work was done and the group dissolved. The subsequent evaluation report identified the following frame factors having a major influence on the project as a whole:

1. The institutional policy:
   a. pressure groups within the institution;
   b. definition and scope of responsibility;
   c. individual initiative;
   d. achievement standards imposed from internal and external bodies;
   e. promotion of faculty members.
2. The political and educational situation:
   a. financial and personnel resources of the development team;
   b. degree of curriculum development in the subject area; access to and availability of instruments of curriculum development;

c. traditional versus innovative school curricula;

d. organizational, material, and educational characteristics of the school.

Apart from the institutional, educational, and political frame factors which conditioned the outcomes of the team's activities, there was a set of others which the team itself could influence and which had a direct influence on the adaptation process:

1. The organization of the activities:
   a. form, division, and control of activities;
   b. sanctions and decision-making;
   c. schedule.
2. The organization of the group:
   a. size of the group;
   b. status and competence difference within the group;
   c. full-time/part-time ratio;
   d. gratification from participation in the adaptation;
   e. contacts with nonparticipants.

The identification of these frame factors led the group to the following findings and conclusions:

1. A scientific institution in its initial phase is continually subject to an internal loss of balance. This instability is due to the low degree of formalized decision-making, vague definition of criteria for success and failure, individual freedom to define work and be subject to examination procedures. and to the high fluctuation of newcomers and leavers.

2. Knowledge of and experience with schools, teachers, and students is a precondition of effective curriculum development. The "ideal" student or teacher is a generalization which can only be made by those who have breached the gap between research and school practice. Curriculum alternatives can only be defined by those who are familiar with the problems as they exist within a school and classroom.

3. Justification of procedures, content, and rationale to a real or fictitious public is often more important than the impact or consequences which the materials have in school. In the case of the SSLU adaptation, the team clung too long to unrealistic aims on the grounds that curriculum development should not sustain inadequate conditions in the social system.

4. As long as project activities are not rewarded in terms of status improvement, they tend to become a matter of minor importance for the team members and continuous and effective work becomes impossible.

Conclusions

1. Participation in project activities should enable team members to qualify in the subject area they have been working in.

2. Project work should be regarded as part of the continuing education for teachers who have been working on the project.

3. Part-time members should be limited in number. They should be relieved from other faculty duties.

4. Project experiences should become part of the general instruction.

5. The project should not cover more than two years.

6. The project should be open to all students.

7. Financial obligations of schools with respect to traditional and innovative materials should be considered.

8. The evaluator should answer to the needs of teachers and make his findings available and intelligible to them. He must insist that the team members themselves develop standards for their working procedures and for curriculum alterations. He must insist on a definition of the rationale underlying the programme and of its goals. There must be a common consent as to acceptable and undesirable procedures. He should, at the beginning of the project, bring about a common consent about schedule and procedure. Disagreement should be discussed openly.

9. Workshops should be held regularly and other project experiences made available.

10. Group dynamics may help to create more consistency among the members.

11. Continuous documentation should make the overall activities transparent.

The role of the evaluator

Neutrality of the evaluator is not desirable. He should share the responsibility for the project. The emphasis of his work and his position within the group should be the result of repeated negotiations with the team members. Neither should the group be authorized to define the evaluator's activities nor should he be authorized to define them all by himself. As the frame factors determine the degree to which group characteristics come uppermost and influence the outcomes of a project, the evaluator should use his influence and knowledge to control these variables and use them to the benefit of the end product. The more he is the professional with a profound knowledge of curriculum theory, evaluation methodology, and school experience, the more his authority is likely to be accepted by his fellow workers.

However discouraging the evaluation report may sound, the benefits from the work on the American materials should not be underestimated. The team had to struggle with the culture-boundedness of the SSLU rationale and develop its own according to a different socio-cultural background. This made them aware of differences and similarities in the political, social, and educational structure of both countries. It sharpened their eye for their own position.

The material in its adapted form was without question an important achievement in the field of West German social-science curricula. The necessity to rearrange the original units, drop parts of them, and add an entirely new one created a different project which required new and heuristic teaching skills.

The open form of the curriculum reflects the rationale of the project: Education towards self-determination is only possible by way of alternatives. The adapted version of the SSLU offered these alternatives to the teacher so that he might use the individual units to the best of a creative interaction with his students. The optimistic vision of the teacher as a friend, a resource person, an organizer of learning processes and model of the enjoying and self-determining individual may or may not be realistic.

The adaptation process itself was beyond doubt a most fruitful exercise in curriculum development for all the team members. The principle of equality, under which they had constituted themselves, was a touchstone for what the project itself was "teaching" and the final completion something like a test for the validity of this idea.

At the moment of this publication the SfK (Sozialwissenschaften für Kinder - Social Science for Children) has not yet been published. Conclusions as to the reception and the effectiveness of the project would be premature. Its potential is promising. Its impact may suffer from undue expectations as to what it can achieve. The SfK aim at long-term attitude changes and should not be judged by immediate student feedback. The project bears a dynamic force. Much depends on the skill of the teacher who uses it. That may be its weakness but that is also its strength.

The evaluation setting was unconventional in that the tasks and the responsibilities of the evaluator were not defined. In the beginning he was tolerated rather then desired. At the end he was urgently needed. He saw himself as a representative of those who were not represented in the adaptation procedure: students and teachers. His struggle for identity and fulfillment was a hard one, but he coped with it and left us with very specific conclusions about his experiences.

REFERENCES/SOURCES

HENTIG, Hartmut von. Einführung in die "Sozialwissenschaften für Kinder". Unpublished manuscript, University of Bielefeld, Fall 1974.

HENTIG, Hartmut von. Personal correspondence. October 1974.

NEUL, Wolfgang. Interaktionsprozesse bei einer Curriculumevaluation. Unpublished manuscript, University of Bielefeld, September 1973.

NEUL, Wolfgang. Einleitung zum Handbuch - Evaluationskapitel. Unpublished manuscript, University of Bielefeld, September 1974.

NEUL, Wolfgang. Personal correspondence, Fall 1974.

SITTE, Karin. Some information on the adaptation of an American social science curriculum, the "Social Science Laboratory Units". Unpublished manuscript, University of Illinois, August 1974.

III. SUGGESTED READING LIST

Evaluation reports and reviews of studies are educative too. A large number of titles was suggested in the Appendix to the OECD publication Evaluating Educational Programmes. The Need and the Response. In addition, the following short list of evaluation study reports and articles indicates several additional useful items. Each was done reflectively and in detail, so as to be instructive to those interested in methodology and commissioning negotiations:

ALKIN, Marvin C., Kosecoff, Jacqueline, Fitz-Gibbon, Carol, and Seligman, Richard. Evaluation and decision-making: The Title VII experience. CSE Monograph Series in Evaluation, Volume 4. Los Angeles: Center for the Study of Evaluation, UCLA, 1974.

BRAUNER, Charles J. The first probe. The Architectural Review, December 1969, CXLVI(874), 451-456. Also in AERA Monograph Series on Curriculum Evaluation, Volume 7. Chicago: Rand McNally, 1974.

BUSSIS, Anne M. and Chittenden, Edward A. Analysis of an approach to open education. PR-70-13. Princeton, New Jersey: Educational Testing Service, August 1970.

FILEP, Robert and Schramm, Wilbur. The impact of research on utilization of media for educational purposes, NDEA Title VII 1958-1968. New York: Institute for Educational Development, 1970.

HARNQVIST, Kjell. The international study of educational achievement. Göteborg, Sweden: University of Göteborg, 1974. (mimeograph)

HIVELY, Wells, et al. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. CSE Monograph Series in Evaluation, Volume 1. Los Angeles: Center for the Study of Evaluation, UCLA, 1973.

HOUSE, ERNEST R. The politics of education innovation. Berkeley: McCutchan Publishing Corporation, 1974.

LARSSON, Inger. Individualized mathematics teaching. Results from the IMU Project in Sweden. Lund: CWK Gleerup, 1973.

LEVIN, Henry M. Cost effectiveness analysis in evaluation research. In Marcia Guttentag (Ed.), Handbook of Evaluation Research. Society for the Psychological Study of Social Issues (in press).

MCLEAN, Leslie D. An evaluation of man and his forests. Toronto, Ontario: Institute for Studies in Education, 1972.

SHIPMAN, M. D., Bolam, D., and Jenkins, D. Inside a curriculum project: A case study in the process of curriculum change. London: Methuen & Co. Ltd., 1974.

SIMONS, Helen. Innovation and the case study of schools. Cambridge Journal of Education, 1971, 1, 118-124.

SKAGER, Rodney. Evaluating educational alternatives. Los Angeles, California: Center for the Study of Evaluation, UCLA, 1974.

SMITH, E. R. and Tyler, R. W. Appraising and recording student progress. New York: Harper & Row, 1942.

SMOCK, H. Richard and Terwilliger, Edith R. An evaluation: Unit 1 - a living-learning experiment at the University of Illinois, Urbana campus, 1971-1972. Urbana, Illinois: Office of Instructional Resources, University of Illinois, 1972. (Mimeograph).